RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI, TIZI-OUZOU

FACULTÉ DES SCIENCES

DÉPARTEMENT DE MATHÉMATIQUES

MÉMOIRE DE MASTER

SPÉCIALITÉ : MATHÉMATIQUES

OPTION : Probabilités et Statistique

Présenté par AMER SAID KENZA

THÈME

Estimation des paramètres d'un modèle de mélange

Devant le jury d'examen composé de :

Mme Abdouche Safia	M.A.A.	UMMTO	Présidente
Mme Ait Mohammed Noura	M.C.B.	UMMTO	Rapporteuse
Mme Merabet Dalila	M.C.B.	UMMTO	Examinatrice

Année universitaire : 2023/2024

Dédicace

Je dédie ce modeste travail :

À ma très chère maman, pour son amour inconditionnel, ses encouragements constants, ses sacrifices inestimables et ses prières qui m'ont toujours accompagnée.

Au meilleur des pères, pour son soutien indéfectible, son affection sincère et la confiance qu'il m'a toujours accordée.

Quoique je puisse dire, je ne saurais exprimer pleinement ma grande affection et ma profonde reconnaissance. J'espère être à la hauteur de leurs attentes.

À mes précieuses sœurs, Kahina et Chahinez, qui sont ma source de force et d'amour.

À mes chers petits frères, Mouloud et Islam, qui apportent tant de bonheur dans ma vie.

À mes adorables cousines, Meriama, ainsi qu'aux petits anges Sarah et Elina.

À mes aimables amies, Sabah, Sarah et Dalia.

À toute ma famille, oncles et tantes, ainsi qu'à toutes les personnes qui m'ont encouragé et soutenu.

Kenza

Remerciements

Je commence par rendre grâce à Dieu, source de toute sagesse et de toute force, pour m'avoir accompagné tout au long de ce parcours. C'est par sa grâce que j'ai pu surmonter les obstacles et trouver l'inspiration nécessaire à la rédaction de ce mémoire.

Je tiens à exprimer ma profonde reconnaissance à ma promotrice Madame **Ait mohammed Noura**, dont l'expertise et la bienveillance ont été des atouts précieux. Ses conseils éclairés et son soutien constant m'ont permis de développer mes idées et d'approfondir mes réflexions.

J'aimerais aussi remercier Madame **Abdouche Safia** d'avoir présidé le jury et Madame **Merabet Dalila** d'avoir examiné ce travail.

Je souhaite également remercier mes enseignants du **Département de Mathématiques de l'Université Mouloud Mammeri de Tizi-Ouzou**, pour leur engagement et leur dévouement. Leur passion pour l'enseignement a nourri ma curiosité et m'a encouragé à donner le meilleur de moi-même.

Je n'oublie pas mes amies et collègues, qui ont partagé avec moi cette aventure. Leurs encouragements et nos échanges enrichissants ont été essentiels à la réalisation de ce travail.

Enfin, je souhaite adresser mes remerciements les plus sincères à ma famille, qui a toujours été mon pilier. Leur soutien inconditionnel, leurs sacrifices et leur amour m'ont permis de poursuivre mes rêves avec détermination.

Table des matières

Table des figures

Liste des tableaux

Liste des symboles et abréviations

Liste des abréviations

i.i.d	Indépendantes et identiquement distribuées
v.a	Variables aléatoires
MV	Maximum Vraisemblance ; de l'anglais Maximum Likelihood
EM	Espérance et maximisation, de l'anglais Expectation Maximization
Θ	Ensemble des paramètres
$\mathcal{F}$	Famille des lois sur Θ
$\mathbb{R}$	Ensembles des réels
$\theta, \mu_1, \mu_2, \sigma_1\sigma_2, \lambda_1, \lambda_2$	Paramètres inconnus
$\widehat{\mu_1}, \widehat{\mu_2}, \widehat{\sigma_1}, \widehat{\sigma_2}, \widehat{\lambda_1}, \widehat{\lambda_2}$	estimateurs de $\mu_1, \mu_2, \sigma_1, \sigma_2, \lambda_1$, et λ_2
π	Distribution de probabilité
$\mathbb{E}$	Espérance mathématique
MCMC	Monte-Carlo par chaînes de Markov, de l'anglais Markov Chain Monte Carlo
$\mathbf{I}_{\mathbb{R}^2}$	Matrice identité

Introduction générale

La classification est importante dans l'analyse des données exploratoires et elle est liée à la décision. Son but est d'évaluer l'homogénéité d'un ensemble d'objets et, en cas de déséquilibre, de les regrouper en sous-ensembles homogènes appelés classes, d'où le terme «classification».
Cependant, la définition de l'homogénéité n'est pas toujours claire, ce qui rend difficile l'établissement de ces classes. Il est donc important de disposer d'un outil adapté au type de données examinées.

Les modèles de mélange sont aujourd'hui largement utilisés en classification et proposent une approche probabiliste. Cette approche considère les objets à classer comme des échantillons de vecteurs aléatoires et met l'accent sur l'analyse de la densité des lois de mélange comme élément central de cette approche.

L'estimation des paramètres dans les modèles de mélange revêt une importance cruciale dans de nombreux domaines tels que la statistique, l'apprentissage automatique et la modélisation probabiliste. L'algorithme EM, dont « EM » signifie Expectation and Maximisation, est un algorithme conçu spécialement pour l'optimisation d'une fonction de vraisemblance, voir par exemple (Dempster et al. [1977]), s'est imposé comme un outil incontournable pour aborder cette problématique complexe. En effet, l'algorithme EM offre une approche itérative et efficace pour estimer les paramètres d'un modèle de mélange, en particulier lorsque les données sont incomplètes ou quand le modèle comporte des variables cachées.

Ce mémoire se propose d'explorer en profondeur l'estimation des paramètres d'un modèle de mélange en se concentrant sur l'utilisation de l'algorithme EM, et sa mise en œuvre pratique. De plus, nous aborderons les principes fondamentaux de l'algorithme EM. Ce dernier est un processus itératif qui utilise la distribution des données complètes pour calculer les estimateurs du maximum de vraisemblance lorsque les données observées sont incomplètes. Ce processus se déroule en deux étapes. La première étape, appelée étape E, consiste à calculer l'espérance conditionnelle de la fonction de log-vraisemblance des données complètes connaissant les données observées. La deuxième étape, l'étape M, implique la maximisation de la log-vraisemblance obtenue à l'étape E pour trouver l'estimateur maximisant cette équation. Ces étapes sont répétées de manière itérative jusqu'à convergence, dans l'espoir d'obtenir l'estimateur du maximum de vraisemblance (voir Kwon [2020] et Haugh [2015]).

Cette étude détaillée démontre l'importance et la pertinence de l'algorithme EM dans le contexte de l'estimation des paramètres des modèles de mélange, et met en évidence ses avantages.

Ce mémoire est articulé en trois chapitres, contribuant de manière significative à la compréhension approfondie de notre sujet.

Le premier chapitre porte sur les définitions clés et les lois usuelles qui serviront à notre analyse tout au long de ce mémoire.

le chapitre 2 se concentre sur l'estimation des paramètres de mélanges plus précisément les mélanges gaussiens, en discutant les modèles utilisés et les difficultés de la méthode du maximum de vraisemblance.

Dans le chapitre trois, nous nous concentrons sur l'approche bayésienne des modèles de mélange, en l'explorant de manière succincte, accompagnée d'un exemple d'application portant sur un modèle gaussien.

Dans le quatrième chapitre, nous effectuons une étude de simulation pour estimer les paramètres d'un modèle de mélange composé de deux distributions gaussiennes. Ce mémoire se termine par une conclusion générale et quelques perspectives.

Chapitre 1

Les modèles de mélanges

1.1 Introduction

Ce chapitre traite les modèles de mélanges, ainsi que les outils statistiques utilisés pour modéliser des données complexes et hétérogènes. Nous donnons aussi quelques lois de probabilités et rappels mathématiques qu'on va utiliser dans ce mémoire.

1.2 Présentation générale

1.2.1 Historique

Les modèles de mélanges ont une origine historique diversifiée, avec des contributions significatives de plusieurs scientifiques éminents. Siméon Denis Poisson a marqué le 19ème siècle en apportant des avancées majeurs dans les domaines des statistiques et des probabilité notamment en établissant la célèbre loi de Poisson. De son coté, Francis Galton s'est penché sur les mélanges de lois binomiales pour étudier la transmission des caractères héréditaires (Korbaa [2006]), tandis que Karl Pearson a également joué un rôle crucial en explorant ces mélanges de lois binomiales.

1.2.2 Définition de modèle de mélange

Le modèle de mélange fini de lois de probabilité suppose que les données proviennent d'une source contenant plusieurs sous-populations homogènes appelées composantes.
La population totale est un assemblage de ces sous-populations, donnant ainsi naissance à un modèle de mélange fini voir Biernacki [2009].

Soit $X = (X_1, ..., X_n)$ un échantillon de variables aléatoires indépendantes identiquement distribués (i.i.d) de loi de mélange fini à K composantes, de densité f dont la forme générale est :

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x), \tag{1.1}$$

avec : π_k les proportions des sous-population telles que $0 \leq \pi_k \leq 1$ et $\sum_{k=1}^{K} \pi_k = 1$, f_k étant la densité du $k^{\text{ème}}$ composante (la paramétrisation des densités des composantes dépend de la nature continue ou discrète des données observées).

Le modèle de mélange est un modèle à données manquantes. La loi de mélange f peut aussi s'interpréter comme la loi marginale de la v.a X_i obtenue a partir de la loi du couple de v.a (X_i, Z_i), $P(X_i = x_i, Z_i = k)$, où $X_i = x_i$ représente la mesure faite sur le ième individu et $Z_i = k$ indique le numéro de la sous-population à laquelle appartient cet individu.

De plus, dans les modèles de mélange les données manquantes sont représentées par $Z = (Z_1, ..., Z_n)$ où $Z_i = k$ indique que l'individu i provient du groupe k.

On constate ainsi que l'échantillon $(X_1, ..., X_n)$ correspond au mélange (1.1) qui peut être interprété comme la loi marginale de la variable X pour le couple (X, Z) (Droesbeke, Saporta et Agnan. [2013]).

1.2.3 Les différents types de modèles de mélange

Les différents types de modèles de mélange sont les suivants :

1. Modèles de mélange paramétriques, qui incluent l'approche de maximum de vraisemblance (MV) souvent utilisées pour estimer les paramètres de ces distributions en utilisant l'algorithme EM (voir par exemple, Danho [2016]).
2. Modèles de mélange semi-paramétriques, combinent des aspects paramétriques et non paramétriques pour offrir une plus grande flexibilité. L'approche bayésienne permet d'incorporer des connaissance a priori sur les paramètres du modèle offrant une estimation plus robuste.
3. Modèles de mélange gaussiens uni-variés, qui sont identifiables. Ces modèles sont souvent utilisés pour la modélisation des données continues.
4. Modèles de mélange gaussiens multidimensionnels, utilisés lorsque les données sont multidimensionnelles. Ils sont couramment utilisés en classification automatique, clustering et modélisation des données complexes basées sur des lois normales multidimensionnelles.

1.3 Lois de probabilités usuelles

1.3.1 La loi gaussienne

La distribution gaussienne, également appelée loi normale, est la distribution la plus familière et couramment utilisée pour modéliser des variables aléatoires continues. Pour une v.a unidimensionnelle X, la fonction de densité de la loi gaussienne est définie par :

$$f(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right), \tag{1.2}$$

avec μ la moyenne, et σ^2 la variance. Si la variable X est d-dimensionnelle , alors la fonction de densité de la loi gaussienne multivariée prend la forme suivante :

$$f(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right). \tag{1.3}$$

Dans cette expression μ représente un vecteur de dimension d, Σ est une matrice de covariance de dimension $d \times d$, et $\det(\Sigma)$ est le déterminant de la matrice Σ.

1.3.2 Loi de Bernoulli $\mathcal{B}(p)$

La loi de Bernoulli, notée $\mathcal{B}(p)$, est une distribution de probabilité définie pour un paramètre $p \in [0;1]$. Elle est caractérisés par une variable aléatoire X prenant ses valeurs dans l'ensemble $\mathcal{V}$={0,1}, avec

$$\mathbb{P}(X = 1) = p \text{ et } \mathbb{P}(X = 0) = 1 - p.$$

Cette loi est couramment utilisée pour modéliser des expériences avec deux issues possibles, telle que échec (0) ou succès (1).

1.3.3 Loi Binomiale $\mathcal{B}(n, p)$

La loi Binomiale, notée $\mathcal{B}(n, p)$, est une distribution de probabilité qui décrit le nombre de succès dans une série de n expériences indépendantes, où chaque expérience a une probabilité de succès p.

Définition 1.1. *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes de loi Bernoulli $\mathcal{B}(p)$. Alors, on dit que la somme $S_n = \sum_{k=1}^{n} X_k$ suit une loi binomiale $\mathcal{B}(n, p)$. Autrement dit, S_n est une variable aléatoire à valeurs dans $\mathcal{V}$={0,...,n}, vérifiant*

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Remarque 1.1. *- La loi binomiale $\mathcal{B}(1, p)$ coïncide avec la loi de Bernoulli $\mathcal{B}(p)$.*
- La loi binomiale $\mathcal{B}(n, p)$ modélise le nombre de succès parmi n expériences indépendantes de Bernoulli.

1.3.4 Loi Beta

Les lois bêta sont des lois de probabilité continues définies sur l'intervalle $]0, 1[$ et paramétrées par deux paramètres de forme, généralement notés α et β. La densité de probabilité

de la loi bêta est définie comme :

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{1.4}$$

où $B(\alpha,\beta)$ est définie par :

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

et $\Gamma(.)$ est la fonction gamma définie par :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,dt.$$

1.3.5 Loi de Dirichlet

La loi de Dirichlet est un outil statistique flexible pour travailler avec des données multinomiales en tenant compte des incertitudes dans les proportions des différentes catégories.

Définition 1.2. *La loi de Dirichlet est une famille de lois de probabilité continues pour des variables aléatoires multinomiales. Elle est paramétrée par un vecteur α de nombres réels positifs et peut être considérée comme une généralisation polynomiale de la loi bêta, et est souvent notée $Dir(\alpha)$.*

Propriétés

- Notation : $Dir(\alpha)$ où $\alpha = (\alpha_1, ..., \alpha_k)$ avec $\alpha_i > 0$ pour $i = 1, ..., k$.
- Densité de probabilité sur le simplex $\{(x_1, ..., x_k) \mid x_i > 0, \Sigma x_i = 1\}$.
- Espérance : $\mathbb{E}(\theta_i) = \alpha_i/\alpha_0$ où $\alpha_0 = \Sigma\alpha_i$, où $\theta_i \sim Dir(\alpha)$.
- Variance : $var(\theta_i) = \alpha_i(\alpha_0 - \alpha_i)/(\alpha_0^2(\alpha_0 + 1))$.

La loi de Dirichlet est une généralisation de la loi bêta qui permet la manipulation de plus de deux catégories. Si $k = 2$, on retrouve la loi bêta.

1.3.6 Loi de Weibull

La distribution de Weibull est une distribution de probabilité continue utilisée pour modéliser la durée de vie et les temps de défaillance, particulièrement dans les domaines de

l'ingénierie de la fiabilité et de l'analyse des risques. Elle est définie par deux paramètres : le **paramètre de forme** k, qui influence la courbe de distribution, et le **paramètre d'échelle** λ, qui détermine l'étendue de la distribution. La fonction de densité de probabilité (PDF) est exprimée par :

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-\left(\frac{t}{\lambda}\right)^k}$$

pour $t \geq 0$, où t représente le temps jusqu'à un événement tel qu'une défaillance. Selon la valeur du paramètre k, la distribution peut modéliser différents comportements : si $k < 1$, le taux de défaillance diminue avec le temps ; si $k = 1$, il reste constant ; et si $k > 1$, il augmente avec le temps. Cette flexibilité permet à la distribution de Weibull d'approcher diverses autres distributions, rendant son utilisation courante dans l'analyse prédictive et la modélisation statistique.

1.4 Quelques rappels mathématiques

1.4.1 Analyse convexe : Fonction convexe

Définition 1.3. *On dit qu'une partie $\mathbb{E}$ du plan $\mathbb{R}^2$ est convexe si pour tout $A, B \in \mathbb{E}$, le segment $[AB]$ est contenu dans $\mathbb{E}$. Une fonction est considérée comme convexe sur un intervalle I si, $\forall x, y \in I$, $\forall t \in [0;1]$,*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{1.5}$$

De plus, une fonction est strictement convexe si,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y). \tag{1.6}$$

Une fonction est concave si son opposé $(-f)$ est convexe, ce qui signifie que

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y). \tag{1.7}$$

Théorème 1.4. *Si la dérivée seconde de f est positive pour tout x dans l'intervalle I où f est deux fois dérivable, alors f est convexe sur cet intervalle.*

Théorème 1.5. *Soit $f : I \to \mathbb{R}$ une fonction convexe, si $(x_1, ..., x_n) \in I$ et si $\lambda_1, ..., \lambda_n \geq 0$, et $\sum_{i=1}^{n} \lambda_i = 1$, alors :*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i). \tag{1.8}$$

Conclusion 1.6. *Les modèles de mélange sont des outils essentiels pour analyser et modéliser des données complexes, et jouent un rôle important dans de nombreux domaines de l'analyse statistique. Dans ce chapitre, nous avons présenté les notions fondamentales sur les modèles de mélanges. Nous avons également présenté quelques lois usuelles de probabilités, telles que la loi de Bernoulli et la loi normale, qui sont souvent utilisées comme composantes dans les modèles de mélange.*

Chapitre 2

Modèles de mélange paramétriques

2.1 Introduction

Dans cette partie, nous mettons en évidence la modélisation des données complexes sous forme de modèles de mélange qui est une combinaison de plusieurs distributions. On estime les paramètres de chaque gaussienne selon le critère du maximum de vraisemblance, qui sera effectué par l'algorithme expectation-maximization (EM).

2.2 Les mélanges paramétriques

Les mélanges paramétriques sont des modèles statistiques qui décrivent la distribution de probabilité d'une variable aléatoire en fonction des paramètres. Les mélanges paramétriques se caractérisent par l'existence d'hypothèses sur la distribution de probabilité induisant une classification. La distribution de probabilité appartient à une famille paramétrique, c'est à dire que l'espace des paramètres est de dimension finie.
En pratique, les modèles de mélange paramétriques sont souvent utilisés en classification automatique pour modéliser des données provenant de différentes sous-populations. Par exemple, dans le cas des mélanges gaussiens, chaque sous-population est distribuée selon une loi gaussienne (voir Melnykov, maitra et al [2010]).

2.2.1 Les types de mélanges paramétriques

Les mélanges paramétriques sont des modèles statistiques qui combinent plusieurs distributions pour représenter des données complexes. Parmi les types les plus courants, on trouve le mélange de gaussiennes, où chaque composante est une distribution gaussienne définie par sa moyenne et sa variance, et le mélange de distributions de Bernoulli, adapté aux données binaires. Les mélanges de distributions de Poisson modélisent le nombre d'événements dans un intervalle fixe, tandis que les mélanges exponentiels sont utilisés pour les temps de survie. Les mélanges de Weibull servent à modéliser les durées de vie ou les temps de défaillance, et les mélanges de Dirichlet sont souvent employés pour les données catégorielles.
Le choix du type de distribution dépend des caractéristiques spécifiques des données à modéliser, la distribution gaussienne étant souvent privilégiée pour sa simplicité et ses propriétés mathématiques.

2.3 L'approche du maximum de vraisemblance : L'algorithme EM

2.3.1 Présentation générale de l'algorithme EM

L'algorithme EM est un algorithme itératif du à Dempster et al. [1977]. Il s'agit d'une méthode d'estimation paramétrique s'inscrivant dans le cadre général du maximum de vraisemblance (MV).

Cet algorithme est largement utilisé dans les modèles de mélange paramétriques, car il permet d'estimer efficacement les paramètres en présence de données incomplètes ou de variables latentes. De manière grossière et vague, il vise à fournir un estimateur lorsque les seules données dont on dispose ne permettent pas l'estimation des paramètres, et que l'expression de la vraisemblance est analytiquement impossible a maximiser.

En résumé, l'algorithme EM procède selon un mécanisme naturel : si un obstacle empêche l'application de la méthode MV, on applique effectivement l'algorithme EM.

2.3.2 Estimation par maximum de vraisemblance

Avant de définir l'algorithme et ses deux étapes, nous allons tout d'abord rappeler le principe de l'estimation par maximum de vraisemblance.

Soit X une variable aléatoire réelle, de loi $\mathfrak{D}_\theta$, de paramètre θ inconnu. On définit une fonction de densité f selon que la loi est discrète ou continue.

- Si X est une variable discrète, alors on pose $f(x;\theta) = P_\theta(X = x)$, c'est à dire la probabilité que X vaut x.

- Si X est une variable continue, alors on pose $f(x;\theta) = f_\theta(x)$, la densité de X au point x.

La vraisemblance du paramètre θ, basée sur les observations $(x_1, ..., x_n)$ d'un échantillon de taille n de v.a indépendantes et identiquement distribuée (i.i.d) selon la loi $\mathfrak{D}_\theta$, est définie comme :

$$L_n(x_1, ..., x_n; \theta) = f(x_1; \theta) \times ... \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta). \tag{2.1}$$

Dans le cas d'une loi discrète, on a :

$$L_n(x_1, ..., x_n; \theta) = \mathbb{P}_\theta\{X_1 = x_1, ..., X_n = x_n\} = \prod_{i=1}^{n} \mathbb{P}_\theta\{X_i = x_i\}. \tag{2.2}$$

L'estimateur MV du paramètre θ est défini comme :

$$\widehat{\theta} = \arg\max_{\theta} L_n(x_1, ..., x_n; \theta) \tag{2.3}$$

Ainsi, pour un échantillon donné, le paramètre θ est choisi de manière à rendre les observations aussi plausibles que possible. Par souci de commodité analytique, il est souvent préférable de maximiser $\log(L_n)$ plutôt que L_n . Étant donné que la fonction logarithme est strictement croissante, cela équivaut au même résultat tout en simplifiant considérablement les calculs :

$$\log(L_n(x_1, ..., x_n; \theta) = \sum_{i=1}^{n} \log f(x_i; \theta). \tag{2.4}$$

Exemple 2.1. *Soit un échantillon observé $x_1, ..., x_n$ généré à partir d'une v.a de Bernoulli de paramètre p.*

$$\mathbb{P}(X_i = x_i; p) = p^{x_i}(1-p)^{(1-x_i)}, x_i \in \{0, 1\}$$

La vraisemblance de p est :

$$L_n(x_1, ..., x_n; p) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)}$$

La log-vraisemblance est donc :

$$l_n(x_i, ..., x_n, p) = \log(L_n(x_1, ..., x_n; p)) = \sum_{i=1}^{n} x_i \log(p) + (n - \sum_{i=1}^{n} x_i) \log(1-p).$$

Calculons la dérivée première de la log-vraisemblance par rapport au paramètre p :

$$\frac{\partial l_n}{\partial p} = \sum_{i=1}^{n}(x_i)\frac{1}{p} - (n - \sum_{i=1}^{n} x_i)\frac{1}{(1-p)}$$

Cette dérivée nous permettra de trouver la valeur de p qui maximise la log-vraisemblance.

Calculons la dérivée seconde de la log-vraisemblance par rapport au paramètre p :

$$\frac{\partial^2 l_n}{\partial p^2} = -\sum_{i=1}^{n}(x_i)\frac{1}{p^2} - (n - \sum_{i=1}^{n} x_i)\frac{1}{(1-p)^2}$$

Cette dérivée nous permettra de vérifier si la fonction de log-vraisemblance est concave, ce qui est une condition nécessaire pour que le maximum de vraisemblance soit atteint.

En annulant la dérivée première de la log-vraisemblance par rapport au paramètre p :

$$\frac{\partial l_n}{\partial p} = 0 \Longleftrightarrow \widehat{p} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

De plus, en vérifiant que la dérivée seconde de la log-vraisemblance par rapport à p est négative

$$\frac{\partial^2 l_n}{\partial p^2} = -\sum_{i=1}^{n}(x_i)\frac{1}{p^2} - (n - \sum_{i=1}^{n} x_i)\frac{1}{(1-p)^2} < 0$$

La dérivée seconde est négative, ce qui signifie que la fonction de log-vraisemblance est concave et atteint son maximum en $\widehat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$. *Cela confirme que l'estimateur du maximum de vraisemblance pour le paramètre p est bien la moyenne des observations de l'échantillon.*

2.3.3 Vraisemblance d'un modèle de mélange

En s'appuyant sur les informations fournies dans le paragraphe précédent, nous pouvons procéder au calcul de la log-vraisemblance d'un modèle de mélange, en tenant compte de l'indépendance des variables X_i :

$$L_n(x_1, ..., x_n, \Theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i; \alpha_k), \tag{2.5}$$

et sa forme logarithmique s'exprime comme suit :

$$l_n(x_1, ..., x_n, \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k(x_i; \alpha_k) \right), \tag{2.6}$$

où $\Theta = (\pi_1, ..., \pi_K, \alpha_1, ..., \alpha_K)$ désigne le vecteur des paramètres inconnus du modèle paramétrique, et $f_k(., \alpha_k)$ représente la densité de la $k^{\text{ème}}$ composante du modèle de mélange paramétrique. Cette version utilise un langage plus fluide et formel tout en conservant la précision des concepts mathématiques.

2.3.4 Algorithme Expectation-Maximization

L'algorithme EM tire son nom du fait qu'à chaque itération, il effectue deux étapes distinctes :

1. La phase **Expectation**, souvent appelée étape **"E"**, consiste à estimer les données inconnues en se basant sur les données observées et les paramètres déterminés lors de l'itération précédente.
2. La phase **Maximization**, ou étape **"M"**, vise à maximiser la vraisemblance en utilisant les estimations des données inconnues obtenues lors de l'étape précédente, et met à jour la valeur du ou des paramètre(s) pour l'itération suivante.

L'algorithme garantit une augmentation de la vraisemblance à chaque itération, ce qui se traduit par l'obtention d'estimateurs de plus en plus précis. Une bonne introduction à l'algorithme EM est donnée par Dempster et al [1977].

Nous allons maintenant définir précisément une itération de l'algorithme, en considérant le cas discret :

- Nous disposons d'observations i.i.d. $X = (x_1, x_2, ..., x_n)$ avec une vraisemblance notée $\mathbb{P}(X \mid \theta)$, dont la maximisation de $\log \mathbb{P}(X \mid \theta)$ est impossible.
- Nous considérons des données cachées $Z = (Z_1, Z_2, ..., Z_n)$ dont la connaissance permettrait de maximiser la vraisemblance des données complètes, $\mathbb{P}(X, Z \mid \theta)$. Puisque les données Z sont inconnues, nous évaluons la vraisemblance des données complètes en tenant compte de toutes les informations disponibles. Pour ce faire, nous utilisons comme estimateur :

$$\mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = \mathbb{E}_{Z|X,\Theta^m}\left[\log \mathbb{P}(X, Z \mid \theta)\right] \quad \text{(étape 'E' de l'algorithme).} \tag{2.7}$$

Pour maximiser cette vraisemblance estimée et obtenir la nouvelle valeur du paramètre (étape 'M' de l'algorithme), nous calculons Θ^{m+1} en fonction de Θ^m :

$$\Theta^{m+1} = arg\max_{\Theta} \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = \text{argmax}\left\{\mathbb{E}_{Z|X,\Theta^m}[\log \mathbb{P}(X, Z \mid \theta)]\right\}. \tag{2.8}$$

2.4 Application de l'algorithme EM aux mélanges paramétriques

2.4.1 Études des mélanges gaussiens

Les mélanges gaussiens, qui supposent que les lois des sous-populations sont distribuées selon une loi normale, suscitent un intérêt important en raison de leur capacité à modéliser des données complexes provenant de diverses sous-populations homogènes Paul et McNicholas. [2016].

Soit le modèle $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$, qui repose sur l'idée que la population étudiée se compose de K sous-populations. Chacune peut être décrite par une densité spécifique $f_k(x)$. Le choix de chaque densité est caractérisée par une proportion π_k, un vecteur de moyenne

μ_k de dimension $b \times 1$, et une matrice de variance-covariance Σ_k de dimension $b \times b$ (voir McNicholas et Peel [2000]).

Pour formaliser cela, il suffit juste de remplacer dans (1.1), $f_k(x)$ par la fonction de densité de la loi gaussienne :

$$f(x \mid \Theta) = \sum_{k=1}^{K} \pi_k f_k(x \mid \theta_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k). \tag{2.9}$$

$\mathcal{N}(x \mid \mu_k, \Sigma_k)$ est la densité de la $k^{\text{ème}}$ composante, et $\Theta = \{(\pi_k, \mu_k, \Sigma_k), k = 1, ..., K\}$.

Étape "E"

Pour pouvoir obtenir les probabilités que les données de chaque individu soient générées par les densités f_k du mélange, il est nécessaire d'estimer les paramètres (π_k,μ_k,Σ_k) de chaque densité. On suppose l'existence d'une matrice Z de dimension $n \times K$, où $z_{ik} = 1$ si l'individu x_i est associé à la densité f_k et z_{ik}= 0 sinon. Les données sont complètes si à la fois X et Z sont observables, mais en pratique, seule X est disponible, rendant les z_{ik} inconnus et les données observées incomplètes. La log-vraisemblance des données est alors calculée pour estimer les paramètres du mélange de densité (Nefkha-Bahri [2020]) :

$$L_n(X, Z \mid \Theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \pi_k f_k(x_i \mid \theta_k), \tag{2.10}$$

ce qui implique

$$\log(L_n(X, Z \mid \Theta)) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(x_i \mid \theta_k)) \tag{2.11}$$

Soit Θ^m les estimations actuelles des paramètres, elles sont des constantes connues, tout comme les données X. Les futures estimations des paramètres et l'ensemble Z sont inconnus. L'algorithme EM est itératif, utilisant les estimations actuelles Θ^m et l'inconnu Z pour calculer les probabilités qu'une observation soit générée par les différentes densités du modèle.

On définit :

$$Q(\Theta \mid \Theta^m) = \mathbb{E}_{Z|X,\Theta^m}[\log(L_n(X, Z \mid \Theta))].$$

Pour chaque k, z_{ik} peut être vu comme une variable de Bernoulli. On a alors :

$$\begin{aligned} Q(\Theta \mid \Theta^m) = &\sum_{i=1}^{n}\sum_{k=1}^{K} \log(\pi_k) P(z_{ik} = 1 \mid X_i = x_i, \Theta^m) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} \log(f_k(x_i \mid \theta_k)) P(z_{ik} = 1 \mid X_i = x_i, \Theta^m), \end{aligned} \tag{2.12}$$

avec la probabilité conditionnelle que l'individu x_i ait été généré par f_k (voir Nefkha-Bahri [2020]) :

$$P_{i,k} = P(z_{ik} = 1 \mid X_i = x_i, \Theta^m) = \frac{\pi_k^{(m)} \det(\Sigma_k^{(m)})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_k^{(m)})^T \left(\Sigma_k^{(m)}\right)^{-1} (x_i - \mu_k^{(m)})\right\}}{\sum_{k'=1}^{K} \pi_{k'}^{(m)} \det(\Sigma_{k'}^{(m)})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_{k'}^{(m)})^T \left(\Sigma_{k'}^{(m)}\right)^{-1} (x_i - \mu_{k'}^{(m)})\right\}} \tag{2.13}$$

Dans ce contexte, $\left(\pi_k^{(m)}, \mu_k^{(m)}, \Sigma_k^{(m)}\right)$, la lettre m située au-dessus des paramètres désigne l'estimation actuelle de ces paramètres. Lorsque x_i est généré par une densité parmi les densités constituant le mélange gaussien, la somme des probabilités est égale à 1 : $\sum_{k=1}^{K} P_{i,k} = 1$.

Nous prenons la fonction **Q** comme un point de départ, des estimateurs peuvent être construits en utilisant la méthode du maximum de vraisemblance (voir McNicholas [2016], p. 335).

Étape "M"

1. **Estimateur de π_k**

   Pour calculer l'estimateur de π_k , il faut satisfaire la contrainte $\sum_{k=1}^{K} \pi_k = 1$. Nous introduisons un multiplicateur de Lagrange λ. Nous devons donc résoudre l'équation

suivante :

$$\frac{\partial}{\partial \pi_k}\left[Q(\Theta \mid \Theta^m) + \lambda(\sum_{k=1}^{K}\pi_k - 1)\right] = 0. \tag{2.14}$$

Pour un certain k,

$$\begin{aligned}\frac{\partial}{\partial \pi_k}\left[Q(\Theta \mid \Theta^m) + \lambda(\sum_{k=1}^{K}\pi_k - 1)\right] &= \frac{\partial}{\partial \pi_k}\left[\sum_{i=1}^{n}\log(\pi_k)P_{i,k} + \lambda(\sum_{k=1}^{K}\pi_k - 1)\right] \\ &= \sum_{i=1}^{n}\frac{1}{\pi_k}P_{i,k} + \lambda = 0.\end{aligned}$$

Cela nous donne $\frac{1}{\pi_k}\sum_{i=1}^{n}P_{i,k} = -\lambda$, donc

$$\pi_k = -\frac{1}{\lambda}\sum_{i=1}^{n}P_{i,k}. \tag{2.15}$$

En effectuant la sommation pour (2.15), on obtient :

$$\begin{aligned}\sum_{k=1}^{K}\pi_k &= -\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=1}^{n}P_{i,k} \\ 1 &= -\frac{1}{\lambda}\sum_{i=1}^{n}(\sum_{k=1}^{K}P_{i,k}) \\ 1 &= -\frac{1}{\lambda}\sum_{i=1}^{n}1 \\ \lambda &= -\sum_{i=1}^{n}1 \\ \lambda &= -n\end{aligned}$$

Ce qui nous donne l'estimateur suivant :

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}P(z_{ik} = 1 \mid X_i = x_i, \Theta^m) = \frac{1}{n}\sum_{i=1}^{n}P_{i,k}. \tag{2.16}$$

2. **Estimateur de μ_k**

Dans le contexte des mélanges gaussiens θ_k=(μ_k, Σ_k) représente les paramètres de la densité f_k, où μ_k est le vecteur moyen et Σ_k est la matrice covariance symétrique.

Par conséquent,

$$\log(f_k(x_i \mid \theta_k)) = C - \frac{1}{2}\log(\det(\Sigma_k)) - \frac{1}{2}(x_i - \mu_k)^T\Sigma_k^{-1}(x_i - \mu_k),$$

avec C une constante sans impact sur les développements futurs. La symétrie de Σ_k nous permet d'écrire, pour un k donné :

$$\begin{aligned}\frac{\partial}{\partial\mu_k}\log(f_k(x_i \mid \theta_k)) &= -\frac{1}{2}\frac{\partial}{\partial\mu_k}\left[(x_i - \mu_k)^T\Sigma_k^{-1}(x_i - \mu_k)\right] \\ &= -\frac{1}{2}\left[(x_i - \mu_k)^T\Sigma_k^{-1}(-1) + (x_i - \mu_k)^T\Sigma_k^{-1}(-1)\right] \\ &= (\Sigma_k^{-1})(x_i - \mu_k). \end{aligned} \tag{2.17}$$

Ainsi, en dérivant la fonction **Q** par rapport à μ_k et égalant à 0, on obtient : $\frac{\partial}{\partial\mu_k}(Q) = \sum_{i=1}^{n}(x_i - \mu_k)^T\Sigma_k^{-1}P_{i,k} = 0$, ce qui permet d'obtenir l'estimateur de μ_k comme suit :

$$\mu_k^{(m+1)} = \frac{\sum_{i=1}^{n} x_i P(z_{ik} = 1 \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(z_{ik} = 1 \mid X_i = x_i, \Theta^m)} = \frac{\sum_{i=1}^{n} x_i P_{i,k}}{\sum_{i=1}^{n} P_{i,k}} \tag{2.18}$$

3. **Estimateur de Σ_k**

Tout d'abord, notons que $\det(\Sigma^{-1}) = \det(\Sigma)^{-1}$. On note la trace et la diagonale de la matrice A respectivement $tr(A)$ et $diag(A)$. Les deux résultats suivants sont obtenus pour deux matrices symétriques A et B (voir Nefkha-Bahri [2020]) :

$$\frac{\partial\log(\det(A))}{\partial A} = 2A - diag(A^{-1});$$
$$\frac{\partial tr(AB)}{\partial A} = B + B^T - diag(B)$$

On considérant que ;

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\log(f_k(x_i \mid \theta_k))P_{i,k} = \sum_{k=1}^{K}\frac{1}{2}\log(det(\Sigma_k^{-1}))\sum_{i=1}^{n}P_{i,k} - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}P_{i,k}tr\left[\Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\right] \tag{2.19}$$

où Σ_k est une matrice symétrique.

Pour un k donné,

$$\begin{aligned}\frac{\partial}{\partial\Sigma_k^{-1}}(Q) &= \frac{1}{2}\sum_{i=1}^{n}P_{i,k}\left(2\Sigma_k - diag(\Sigma_k)\right) - \frac{1}{2}\sum_{i=1}^{n}P_{i,k}\Big(2(x_i-\mu_k)(x_i-\mu_k)^T \\ &\quad - diag\left[(x_i-\mu_k)(x_i-\mu_k)^T\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}P_{i,k}(2[\Sigma_k-(x_i-\mu_k)(x_i-\mu_k)^T] - diag[\Sigma_k-(x_i-\mu_k)(x_i-\mu_k)^T]) \\ &= 0\end{aligned}$$

Il s'agit d'une représentation matricielle sous la forme 2D - diag$(D) = 0$, cela signifie que D = 0 puisque 2D = diag(D).

Donc,

$$\sum_{i=1}^{n}P_{i,k}\Big(2[\Sigma_k-(x_i-\mu_k)(x_i-\mu_k)^T]\Big) = 0.$$

Cela conduit à l'estimateur suivant :

$$\Sigma_k^{(m+1)} = \frac{\sum_{i=1}^{n}P_{i,k}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}P_{i,k}} \tag{2.20}$$

L'algorithme s'arrête après un nombre prédéfini d'itérations ou bien à la stationnarité du critère de log-vraisemblance observée. Cette seconde option découle de la propriété de la croissance de $L_n(\theta; x)$ à chaque itération.

Algorithme EM Standard

Entrées :

— $\Theta^{(0)}$: une valeur initiale des paramètres du modèle

— X : un ensemble d'observations

— C_j : nombre de classes

— ϵ : un seuil de convergence pour l'algorithme

Sorties : La valeur de Θ qui maximise la vraisemblance.

1. Initialiser m à 0.
2. Répéter jusqu'à convergence :
   (a) **Étape E** : Calculer l'espérance conditionnelle de la fonction de vraisemblance :

$$Q(\Theta|\Theta^{(m)}) = \mathbb{E}_{Z|X,\Theta^m}[\log(L_n(X, Z \mid \Theta))]$$

   Calcule de probabilité conditionnelle :

$$P_{i,k} = P(z_{ik} = 1 \mid X_i = x_i, \Theta^m) = \frac{\pi_k^{(m)} f_k(x_i)}{\sum_{k'=1}^{K} \pi_{k'}^{(m)} f_{k'}(x_i)}$$

   (b) **Étape M** : Maximiser $Q(\Theta|\Theta^{(m)})$:

$$\Theta^{(m+1)} = \arg\max_{\theta} Q(\Theta; \Theta^{(m)})$$

   (c) Incrémenter m de 1.
3. Arrêter lorsque :

$$Q(\Theta|\Theta^{(m+1)}) - Q(\Theta|\Theta^{(m)}) < \epsilon$$

4. A la convergence de l'algorithme, on peut déduire une partition en rangeant chaque individu dans la classe maximisant la probabilité conditionnelle $P_{i,k}$.

2.5 Exemples de mélange de deux gaussiennes

Exemple 2.2. *Considérons $X = (X_1, ..., X_n)$ comme un échantillon i.i.d d'observations provenant d'un mélange de deux gaussiennes unidimensionnelles, et $Z = (Z_1, ..., Z_n)$ comme les données cachées où Z_i indique la distribution dont X_i provient :*

Lorsque X_i provient de la distribution 1 : $L(X_i \mid Z_i = 1) = \mathcal{N}_1(\mu_1, \sigma_1^2)$

Lorsque X_i provient de la distribution 2 : $L(X_i \mid Z_i = 2) = \mathcal{N}_1(\mu_2, \sigma_2^2)$

avec les probabilités associées : $\mathbb{P}(Z_i = 1) = \pi_1$ *et* $\mathbb{P}(Z_i = 2) = \pi_2 = 1 - \pi_1$

L'objectif est d'estimer les 5 paramètres inconnus $\Theta = (\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

La vraisemblance des données complètes est exprimée par :

$$L_n(X, Z \mid \Theta) = \prod_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \pi_j f_j(x_i)$$

où $f_j : \mathbb{R} \to \mathbb{R}$ *telle que :*

$$f_j(x) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x - \mu_j}{\sigma_j})^2)$$

représente une densité gaussienne unidimensionnelle avec les paramètres μ_j *et* σ_j^2.

La log-vraisemblance des données complètes obtenue est :

$$\begin{aligned}
\log(L_n(X, Z \mid \Theta)) &= \log\left(\prod_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \pi_j f_j(X_i)\right) \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \left(\log(\pi_j) - \frac{1}{2}\log(2\pi) - \log(\sigma_j) - \frac{1}{2}\left(\frac{X_i - \mu_j}{\sigma_j}\right)^2\right)\right] \\
&= \sum_{i=1}^{n} \left[\mathbb{1}_{\{Z_{i=1}\}} \left(\log(\pi_1) - \frac{1}{2}\log(2\pi) - \log(\sigma_1) - \frac{1}{2\sigma_1^2}\left(X_i - \mu_1\right)^2\right)\right. \\
&\quad \left. + \mathbb{1}_{\{Z_{i=2}\}} \left(\log(\pi_2) - \frac{1}{2}\log(2\pi) - \log(\sigma_2) - \frac{1}{2\sigma_2^2}\left(X_i - \mu_2\right)^2\right)\right]
\end{aligned}$$

A chaque itération, l'étape "E" implique la nécessité de définir la distribution de Z_j connaissant X_j et Θ^m, en utilisant la formule de Bayes. Cette distribution conditionnelle est calculée en fonction des paramètres estimés à l'itération précédente. Plus précisément, la distribution de Z_j sachant X_j et Θ^m est obtenue en calculant la probabilité que Z_j prenne une valeur donnée, étant donné les observations X_j et les paramètres du modèle à l'itération courante.

$$P_{i,j} = P(Z_i = j \mid X_i = x_i, \Theta^m) = \frac{\pi_j f_j(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)}$$

Calculons maintenant l'espérance conditionnelle :

$$\mathbb{E}_{Z|X,\Theta^m}\left[\log(L_n(X, Z \mid \Theta))\right],$$

conditionnellement aux données observées X et à la valeur actuelle du vecteur des paramètres Θ.

$$\begin{aligned}\mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) &= \mathbb{E}_{Z|X,\Theta^m}\left[\log(L_n(X, Z \mid \Theta))\right] \\ &= \sum_{i=1}^{n}\left[\sum_{j=1}^{2} P_{i,j}\left(\log(\pi_j) - \frac{1}{2}\log(2\pi) - \log(\sigma_j) - \frac{1}{2}\left(\frac{X_i - \mu_j}{\sigma_j}\right)^2\right)\right] \\ &= \sum_{i=1}^{n}\left[P_{i,1}\left(\log(\pi_1) - \frac{1}{2}\log(2\pi) - \log(\sigma_1) - \frac{1}{2\sigma_1^2}\left(X_i - \mu_1\right)^2\right)\right. \\ &\quad \left. + P_{i,2}\left(\log(\pi_2) - \frac{1}{2}\log(2\pi) - \log(\sigma_2) - \frac{1}{2\sigma_2^2}\left(X_i - \mu_2\right)^2\right)\right]\end{aligned}$$

On cherche a maximiser cette vraisemblance estimée des données complètes pour déterminer les nouvelles valeurs des paramètres $(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

1. $\frac{\partial}{\partial \pi_1}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow \sum_{i=1}^{n} P_{i,1}\frac{1}{\pi_1} - \sum_{i=1}^{n} P_{i,2}\frac{1}{1-\pi_1} = 0$

2. $\frac{\partial}{\partial \pi_2}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow \sum_{i=1}^{n} P_{i,1}\frac{1}{1-\pi_2} - \sum_{i=1}^{n} P_{i,2}\frac{1}{\pi_2} = 0$

3. $\frac{\partial}{\partial \mu_1}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow \sum_{i=1}^{n} P_{i,1}\left[-\left(x_i - \mu_1\right)\right] = -\sum_{i=1}^{n} P_{i,1}x_i + \sum_{i=1}^{n} P_{i,1}\mu_1 = 0$

4. $\frac{\partial}{\partial \mu_2}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow -\sum_{i=1}^{n} P_{i,2}x_i + \sum_{i=1}^{n} P_{i,2}\mu_2 = 0$

5. $\frac{\partial}{\partial \sigma_1}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow -\sum_{i=1}^{n} P_{i,1}\frac{1}{\sigma_1} + \sum_{i=1}^{n} P_{i,1}\left(\frac{2\sigma_1}{2\sigma_1^4}(x_i - \mu_1)\right) = -\sum_{i=1}^{n} P_{i,1}\frac{1}{\sigma_1}$
   $+ \sum_{i=1}^{n} P_{i,1}\left(\frac{1}{\sigma_1^3}(x_i - \mu_1)\right) = 0$

6. $\frac{\partial}{\partial \sigma_2}\ \mathbf{Q}(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{\mathbf{m}}) = 0 \Leftrightarrow -\sum_{i=1}^{n} P_{i,2}\frac{1}{\sigma_2} + \sum_{i=1}^{n} P_{i,2}\left(\frac{2\sigma_2}{2\sigma_2^4}(x_i - \mu_1)\right) = -\sum_{i=1}^{n} P_{i,2}\frac{1}{\sigma_2}$
   $+ \sum_{i=1}^{n} P_{i,2}\left(\frac{1}{\sigma_2^3}(x_i - \mu_2)\right) = 0$

Un calcul rapide aboutit à la prochaine mise à jour $\Theta^{(m+1)}$ des estimateurs déjà identifiés $\Theta^{(m)}$ (pour j = 1 et 2)

1. ***Estimateur de*** π_1 :

   Pour $j = 1$, calculer la somme pondérée des probabilités $P_{i,1}$ sur l'ensemble des données : $\sum_{i=1}^{n} P_{i,1}$.

   Mettre à jour π_1 en divisant cette somme par le nombre total de données n :

$$\pi_1^{(m+1)} = \tfrac{1}{n} \sum_{i=1}^{n} P(Z_i = 1 \mid X_i = x_i, \Theta^m) = \frac{1}{n} \sum_{i=1}^{n} P_{i,1}$$

2. ***Estimateur de*** π_2 ***:***

   Pour $j = 2$, calculer la somme pondérée des probabilités $P_{i,2}$ sur l'ensemble des données : $\sum_{i=1}^{n} P_{i,2}$.

   Mettre à jour π_2 en divisant cette somme par le nombre total de données n :

$$\pi_2^{(m+1)} = \tfrac{1}{n} \sum_{i=1}^{n} P(Z_i = 2 \mid X_i = x_i, \Theta^m) = \frac{1}{n} \sum_{i=1}^{n} P_{i,2}$$

3. ***Estimateur de*** μ_1 ***:***

   Pour $j = 1$, calculer la somme pondérée des observations multipliées par les probabilités associées $P_{i,1}$: $\sum_{i=1}^{n} P_{i,1}$.

   Mettre à jour μ_1 en divisant cette somme des probabilités pour cette composante :

$$\mu_1^{(m+1)} = \frac{\sum_{i=1}^{n} x_i P(Z_i = 1 \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = 1 \mid X_i = x_i, \Theta^m)} = \frac{\sum_{i=1}^{n} x_i P_{i,1}}{\sum_{i=1}^{n} P_{i,1}}$$

4. ***Estimateur de*** μ_2 ***:***

   Pour $j = 2$, calculer la somme pondérée des observations multipliées par les probabilités associées $P_{i,2}$: $\sum_{i=1}^{n} P_{i,2}$.

   Mettre à jour μ_2 en divisant cette somme des probabilités pour cette composante :

$$\mu_2^{(m+1)} = \frac{\sum_{i=1}^{n} x_i P(Z_i = 2 \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = 2 \mid X_i = x_i, \Theta^m)} = \frac{\sum_{i=1}^{n} x_i P_{i,2}}{\sum_{i=1}^{n} P_{i,2}}$$

5. ***Estimateur de*** σ_1^2 ***:***

   Pour $j = 1$, calculer la somme pondérée des carrés des écarts entre les observations et la moyenne au carré, pondérée par les probabilités associées $P_{i,1}$: $\sum_{i=1}^{n} P_{i,1}(x_i - \mu_1^{(m+1)})^2$.

Mettre à jour $\sigma_1^{2(m+1)}$ *en divisant cette somme des probabilités pour cette composante :*

$$\sigma_1^{2(m+1)} = \frac{\sum_{i=1}^{n}(x_i - \mu_1^{(m+1)})^2 P(Z_i = 1 \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = 1 \mid X_i = x_i, \Theta^m)}$$
$$= \frac{\sum_{i=1}^{n} P_{i,1}(x_i - \mu_1^{(m+1)})(x_i - \mu_1^{m+1})^2}{\sum_{i=1}^{n} P_{i,1}}$$

6. ***Estimateur de*** σ_2^2 ***:***

Pour $j = 2$*, calculer la somme pondérée des carrés des écarts entre les observations et la moyenne au carré, pondérée par les probabilités associées* $P_{i,2}$ *:* $\sum_{i=1}^{n} P_{i,2}(x_i - \mu_2^{(m+1)})^2$

Mettre à jour $\sigma_2^{2(m+1)}$ *en divisant cette somme des probabilités pour cette composante :*

$$\sigma_2^{2(m+1)} = \frac{\sum_{i=1}^{n}(x_i - \mu_2^{(m+1)})^2 P(Z_i = 2 \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = 2 \mid X_i = x_i, \Theta^m)}$$
$$= \frac{\sum_{i=1}^{n} P_{i,2}(x_i - \mu_2^{(m+1)})(x_i - \mu_2^{m+1})^2}{\sum_{i=1}^{n} P_{i,2}}$$

A la convergence de l'algorithme, on peut déduire une partition en rangeant chaque individu dans la classe maximisant la probabilité conditionnelle $P_{i,j}$*.*

Exemple 2.3. *On reprend l'exemple précédent mais avec des observations bidimensionnelles*

$$\textit{Lorsque } X_i \textit{ provient de la distribution 1 : } L(X_i \mid Z_i = 1) = \mathcal{N}_2(\mu_1, \Sigma_1)$$
$$\textit{Lorsque } X_i \textit{ provient de la distribution 2 : } L(X_i \mid Z_i = 2) = \mathcal{N}_2(\mu_2, \Sigma_2)$$

La vraisemblance est exprimée par :

$$L_n(X, Z \mid \Theta) = \prod_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \pi_j f_j(X_i)$$

Où $f_j : \mathbb{R}^2 \to \mathbb{R}$ *telle que :*

$$f_j(x) = \mathcal{N}(x \mid \mu_j, \Sigma_j) = \frac{1}{2\pi \det(\Sigma_j)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right) \textit{ (densité gaussienne bidimensionnelle avec les paramètres } (\mu_j, \Sigma_j)).$$

La log-vraisemblance est :

$$\log(L_n(X, Z \mid \Theta) = \log(\prod_{i=1}^{n} \sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \pi_j \mathcal{N}(x \mid \mu_j, \Sigma_j))$$

$$= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} \mathbb{1}_{\{Z_{i=j}\}} \left(\log(\pi_j) - \log(2\pi) - \frac{1}{2}\log(\det(\Sigma_j)) - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) \right) \right] \quad (2.21)$$

A chaque itération, nous devons spécifier la probabilité conditionnelle de Z_j *connaissant* X_i *et* Θ^m*. On définit :*

$$P_{i,j} = P(Z_i = j \mid X_i = x_i, \Theta^m) = \frac{\pi_j f_j(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)}$$

avec $f_j(x_i) \equiv N(x_i \mid \mu_j, \sigma_j^2)$ *; alors :*

$$P_{i,j} = \frac{\pi_j^{(m)} \det(\Sigma_j^{(m)})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \mu_j^{(m)})^T \left(\Sigma_j^{(m)}\right)^{-1} (x_i - \mu_j^{(m)})\}}{\pi_1^{(m)} \det(\Sigma_1^{(m)})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \mu_1^{(m)})^T \left(\Sigma_1^{(m)}\right)^{-1} (x_i - \mu_1^{(m)})\} + \pi_2^{(m)} \det(\Sigma_2^{(m)})^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_i - \mu_2^{(m)})^T \left(\Sigma_2^{(m)}\right)^{-1} (x_i - \mu_2^{(m)})\}}$$

En appliquant (2.8), on trouve :

$$\begin{aligned} Q(\Theta \mid \Theta^m) &= \mathbb{E}_{Z|X,\Theta^m}[\log(L_n(X, Z \mid \Theta))] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} P_{i,j} \log\left(\pi_j f_j(X_i)\right) \\ &= \sum_{i=1}^{n} \left[\sum_{j=1}^{2} P_{i,j} \left(\log(\pi_j) - \log(2\pi) - \frac{1}{2}\log(\det(\Sigma_j)) - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) \right) \right] \end{aligned}$$

Il ne reste plus qu'à réaliser l'étape "M" , qui consiste a maximiser en Θ *l'expression* $\mathbf{Q}(\mathbf{\Theta} \mid \mathbf{\Theta^m})$. *Nous cherchons un triplet* $\Theta_j^{(m+1)} = (\pi_j^{(m+1)}; \mu_j^{(m+1)}, \Sigma_j^{(m+1)})$ *qui annule son gradient, donc tel que :*

1. $\frac{\partial}{\partial \pi_j}\left[\mathbf{Q}(\mathbf{\Theta} \mid \mathbf{\Theta^m}) + \lambda(\sum_{J=1}^{2} \pi_j - 1)\right] = 0 \Leftrightarrow \frac{\partial}{\partial \pi_j}\sum_{i=1}^{n} \pi_j \log(\pi_j) P_{i,j} = \sum_{i=1}^{n} \frac{1}{\pi_j} P_{i,j} = 0$ *(λ le multiplicateur de Lagrange)*
2. $\frac{\partial}{\partial \mu_j}\ (Q) = \sum_{i=1}^{n} (x_i - \mu_j)^T \Sigma_j^{-1} P_{i,j} = 0$
3. $\frac{\partial}{\partial \Sigma_j^{-1}}(Q) = \frac{1}{2}\sum_{i=1}^{n} P_{i,j}\,(2\Sigma_j - diag(\Sigma_j))$ - $\frac{1}{2}\sum_{i=1}^{n} P_{i,j}\left(2(x_i - \mu_j)(x_i - \mu_j)^T - diag\left[(x_i - \mu_j)(x_i - \mu_j)^T\right]\right. = \frac{1}{2}\sum_{i=1}^{n} P_{i,j}\left(2[\Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^T] - diag[\Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^T]\right. = \sum_{i=1}^{n} P_{i,j}\left(2[\Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^T]\right) = 0$

Un calcul rapide aboutit à la prochaine mise à jour $\Theta^{(m+1)}$ *des estimateurs déjà calculés* $\Theta^{(m)}$ *(pour j = 1 ou 2)*

$$\pi_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} P(Z_i = j \mid X_i = x_i, \Theta^m) = \frac{1}{n}\sum_{i=1}^{n} P_{i,j}$$

$$\mu_j^{(m+1)} = \frac{\sum_{i=1}^{n} x_i P(Z_i = j \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = j \mid X_i = x_i, \Theta^m)} = \frac{\sum_{i=1}^{n} x_i P_{i,j}}{\sum_{i=1}^{n} P_{i,j}}$$

$$\Sigma_j^{(m+1)} = \frac{\sum_{i=1}^{n} (x_i - \mu_j^{(m+1)})(x_i - \mu_j^{(m+1)})^T P(Z_i = j \mid X_i = x_i, \Theta^m)}{\sum_{i=1}^{n} P(Z_i = j \mid X_i = x_i, \Theta^m)}$$

$$= \frac{\sum_{i=1}^{n} P_{i,j}(x_i - \mu_j^{(m+1)})(x_i - \mu_j^{m+1})^T}{\sum_{i=1}^{n} P_{i,k}}$$

Remarque 2.1. *-On note que ces formules se généralisent de manière directe aux distributions gaussiennes p-dimensionnelles, pour tout $p > 2$.*
- Il est noté aussi que le seul estimateur qui est conditionné par un autre est Σ_j, qui dépend de l'estimateur μ_j. Ainsi, une séquence logique pour réaliser les estimations lors de l'étape M serait d'estimer d'abord π_j, puis μ_j et enfin Σ_j.

2.6 Croissance de la vraisemblance d'une itération a l'autre

L'objectif est de mettre à jour la valeur de $\Theta^m \in \mathbb{R}^d$ vers une version meilleur Θ, qui améliore la vraisemblance (voir Collins [1997] et Morgenthaler [2008]), telle que

$$\Delta(\Theta, \Theta^m) := \log(L_n(X \mid \Theta)) - \log(L_n(X \mid \Theta^m)) \geq 0.$$

Nous souhaiterions que la différence soit aussi grande que possible . Toutefois, comme mentionné dans la définition d'une itération, nous savons pas comment maximiser $L_n(X \mid \Theta)$, ce qui signifie que nous ne savons même pas comment maximise $\Delta(\Theta, \Theta^m)$. Malgré tout, il existe des moyens de l'optimiser, cette différence peut s'écrire :

$$\begin{cases} \Delta(\Theta, \Theta^m) \geq \delta(\Theta, \Theta^m) & \forall \Theta^m \in \mathbb{R}^d \\ \delta(\Theta^m, \Theta^m) = 0 \end{cases} \tag{2.22}$$

Par conséquent, $\delta(\Theta, \Theta^m)$ sert de borne inférieure pour $\Delta(\Theta, \Theta^m)$, et son maximum est supérieur ou égal à 0. En trouvant un Θ' qui maximise $\Theta \longmapsto \delta(\Theta, \Theta^m)$, nous obtenons automatiquement une valeur plus probable pour Θ', qui est $\Delta(\Theta', \Theta^m) \geq 0$. Pour trouver une telle fonction δ, nous avons recours à une représentation marginale de la vraisemblance fondée sur les variables aléatoires cachées $Z = (Z_1, Z_2, ..., Z_n)$:

$$L_n(X \mid \Theta) = \sum_Z L_n(X, Z \mid \Theta) = \sum_Z L_n(X \mid Z, \Theta) L_n(Z \mid \Theta)$$

On obtient alors l'équation suivante :

$$\begin{aligned}\Delta(\Theta,\Theta^m) &= \log(L_n(X \mid \Theta)) - \log(L_n(X \mid \Theta^m)) \\ &= \log(\sum_Z L_n(X \mid Z,\Theta)L_n(Z \mid \Theta)) - \underbrace{\sum_Z L_n(Z \mid X,\Theta^m)}_{=1} \log\left(L_n(X \mid \Theta^m)\right) \quad (2.23)\end{aligned}$$

Cette formulation implique l'utilisation du logarithme de la somme, et en appliquant l'inégalité de Jensen (1.8), on voit clairement comment réduire $\Delta(\Theta,\Theta^m)$. En réécrivant l'équation (2.22) en incorporant les $L_n(Z \mid X,\Theta^m)$ de la somme de droite dans celle de gauche, on ce rapproche de l'expression (2.8) :

$$\Delta(\Theta,\Theta^m) = \log\left(\sum_Z \frac{L_n(X \mid Z,\Theta)L_n(Z \mid \Theta)}{L_n(Z \mid X,\Theta^m)} L_n(Z \mid X,\Theta^m)\right) - \sum_Z L_n(Z \mid X,\Theta^m)\log\left(L_n(X \mid \Theta^m)\right)$$

En observant que $\sum_Z L_n(Z \mid X,\Theta^m) = 1$, nous utilisons ensuite l'inégalité de Jensen :

$$\Delta(\Theta,\Theta^m) \geq \sum_Z L_n(Z \mid X,\Theta^m)\log\left(\frac{L_n(X \mid Z,\Theta)L_n(Z \mid \Theta)}{L_n(Z \mid X,\Theta^m)}\right) - \sum_Z L_n(Z \mid X,\Theta^m)\log\left(L_n(X \mid \Theta^m)\right).$$

De plus

$$\begin{aligned}&\sum_Z L_n(Z \mid X,\Theta^m)\log\left(\frac{L_n(X \mid Z,\Theta)L_n(Z \mid \Theta)}{L_n(Z \mid X,\Theta^m)}\right) - \sum_Z L_n(Z \mid X,\Theta^m)\log\left(L_n(X \mid \Theta^m)\right) \\ &= \sum_Z L_n(Z \mid X,\Theta^m)\log\left(\frac{L_n(X \mid Z,\Theta)L_n(Z \mid \Theta)}{L_n(Z \mid X,\Theta^m)L_n(X \mid \Theta^m)}\right) \\ &= \sum_Z L_n(Z \mid X,\Theta^m)\log\left(\frac{L_n(X,Z \mid \Theta)}{L_n(X,Z \mid \Theta^m)}\right) \\ &=: \delta(\Theta \mid \Theta^m)\end{aligned}$$

Ainsi, nous avons trouvé une fonction $\Theta \longmapsto \delta(\Theta,\Theta^m)$ qui satisfait les conditions (2.22), il est clair que $\delta(\Theta^m,\Theta^m) = 0$. Finalement , nous posons :

$$\begin{aligned}
\Theta^{m+1} &= \arg\max_{\theta} \delta(\Theta, \Theta^m) \\
&= \arg\max_{\theta} \left\{ \sum_{Z} L_n(Z \mid X, \Theta^m) \log \left(\frac{L_n(X, Z \mid \Theta)}{L_n(X, Z \mid \Theta^m)} \right) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{Z} L_n(Z \mid X, \Theta^m) \log(L_n(X, Z \mid \Theta)) \right\} \\
&= \arg\max_{\theta} \left\{ \mathbb{E}_{Z|X,\Theta^m}[\log L_n(X, Z \mid \Theta)] \right\} \qquad (2.24)
\end{aligned}$$

Ainsi, on identifie une valeur Θ^{m+1} plus probable que Θ^m, car :

$$\log L_n(X \mid \Theta^{m+1}) - \log L_n(X \mid \Theta^m) = \Delta(\Theta^{m+1}, \Theta^m) \geq \delta(\Theta^{m+1} \mid \Theta^m) \geq \delta(\Theta^m \mid \Theta^m) \geq 0.$$

Conclusion 2.4. *Dans ce chapitre, nous avons étudié en détail l'estimation des paramètres d'un modèle de mélange gaussien, une approche statistique puissante pour modéliser des données complexes issues de populations hétérogènes. Nous avons estimé les paramètres des modèles de mélange gaussiens par la méthode du maximum de vraisemblance effectuée par l'algorithme EM.*

Chapitre 3

L'approche bayésienne

3.1 Introduction

Les approches bayésiennes ont une histoire riche et diversifiée qui s'étend sur plusieurs siècles. Le nom vient de Thomas Bayes, un mathématicien du $18^{\text{ème}}$ siècle qui a formulé un exemple spécifique du théorème de Bayes dans son article posthume de 1763. Cette approche repose sur l'utilisation de probabilités conditionnelles pour mettre à jour les croyances initiales à partir de nouvelles données observées. Bien que l'histoire des approches bayésiennes soit pleine de développements et de controverses, elles ont gagné en popularité au cours des dernières décennies en raison de leurs avantages théoriques et pratiques dans de nombreux domaines tels que l'analyse des données, l'intelligence artificielle, la médecine et la finance.

Pour approfondir votre compréhension des concepts et des méthodes de la statistique bayésienne, nous vous suggérons de consulter les références ci-dessous Carlin et louis [1997] , Gelman et al [1995] et Robert [2006].

3.2 Inférence Bayésienne

3.2.1 Théorème de Bayes

Si l'on considère un vecteur $x = (x_1, ..., x_n)'$ représentant n observations provenant d'une distribution $\pi(x \mid \theta)$, et que le paramètre θ suit une distribution de probabilité $\pi(\theta)$, alors

$$\begin{aligned} \pi(x \mid \theta)\pi(\theta) &= \pi(x, \theta) \\ &= \pi(\theta \mid x)\pi(x) \end{aligned} \tag{3.1}$$

La distribution conditionnelle de θ, étant donné les observations $(x_1, ..., x_n)$, est

$$\pi(\theta \mid x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} \tag{3.2}$$

La formule (3.2) est appelée la formule de Bayes. La signification de chaque notation dans cette formule est :

- $\pi(\theta \mid x)$ représente la distribution conditionnelle du paramètre θ sachant les observations $(x_1, ..., x_n)$. C'est la probabilité de θ étant donné les données observées x.

- $\pi(x \mid \theta)$ représente la probabilité des données observées sachant le paramètre θ.

- $\pi(\theta)$ est la distribution a priori du paramètre θ, c'est-à-dire la probabilité de θ avant d'observer les données.

- $\pi(x)$ probabilité marginale des données x, utilisée pour normaliser la distribution conditionnelle et obtenir la distribution a posteriori.

3.2.2 Loi a priori

Définition 3.1. *La loi a priori est une distribution de probabilité qui résume l'ensemble des informations disponibles sur un paramètre d'intérêt, avant même de recueillir les données.*

Cette loi indique la probabilité que le paramètre prenne telle ou telle valeur, en se basant sur des connaissances préalables.

Il existe plusieurs types de lois a priori :
- Lois a priori informatives, qui prennent en compte des informations externes (avis d'experts, expériences antérieures, etc).
- Lois a priori neutres ou non informatives, où toutes les valeurs du paramètre ont la même probabilité a priori.

Choix de la loi a priori

Le choix de la loi a priori est une étape cruciale en statistique bayésienne. Ce choix peut être motivé par diverses considérations, telles que les expériences passées ou une intuition par rapport à un phénomène étudié, les aspects de calculabilité en choisissant une loi conjuguée à la vraisemblance, ou encore l'absence d'information, justifiant l'utilisation d'une loi non informative.

3.2.3 Lois a priori conjuguées

Une loi a priori conjuguée est une famille de lois qui, pour une loi a priori particulière, produit une loi a posteriori appartenant à la même famille. Les lois conjugués présentent plusieurs avantages, notamment un calcul plus facile et la possibilité de mettre à jour rapidement la loi a posteriori dans certains cas.

Définition 3.2. *(Robert [2006])*

Une famille $\mathcal{F}$ de lois sur Θ est dite conjuguée si, pour tout π appartenant à cette famille, la loi $\pi(\theta \mid x)$ appartient également à la même famille $\mathcal{F}$.
En d'autres termes, si on part d'une loi a priori conjuguée et qu'on observe des données x, alors la loi a posteriori mise à jour aura la même forme que la loi a priori, mais avec des paramètres différents.

Les lois a priori conjuguées ne peuvent être dérivées qu'à partir des familles de lois exponentielles, qui se présentent sous la forme suivante :

$$f(x \mid \theta) = C(\theta)h(x)\exp[R(\theta).T(x)].$$

Où :

- C est une fonction des paramètres.
- h est une fonction des données.
- R et T sont des fonctions vectorielles des paramètres et des données respectivement.

Pour ces familles exponentielles, la forme générale des lois a priori conjuguées est (voir par exemple Robert [1996]) :

$$\pi(\theta \mid \nu, \lambda) = K(\nu, \lambda)\exp\{\theta.\nu - \lambda\psi(\theta)\}; \tag{3.3}$$

Où ν et λ sont les paramètres de la loi a priori, et $K(\nu, \lambda)$ est une constante de normalisation qui assure que π est une densité de probabilité valide.

Avantages des Lois Conjuguées

L'avantage des lois conjuguées est que les calculs de la loi a posteriori sont simplifiés, car elle appartient à la même famille que la loi a priori, avec des paramètres mis à jour en fonction des données observées. Cela permet d'obtenir des expressions analytiques pour l'inférence bayésienne dans certains cas.

La loi a posteriori se présente généralement sous la forme $\pi(\theta|\nu+x, \lambda+1)$, où x représente les données observées. Des exemples de lois conjuguées courantes peuvent être trouvés dans l'ouvrage de Robert et Casella [1999].

Familles de lois a priori conjuguées : quelques illustrations

$f(x \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid x)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
$\mathcal{P}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + x, \beta + 1)$
$\Gamma(\nu, \theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + \nu, \beta + x)$
$\mathcal{B}(n, \theta)$	$Be(\alpha, \beta)$	$Be(\alpha + x, \beta + n - x)$
$\mathcal{M}_k(\theta_1, ..., \theta_k)$	$\mathcal{D}(\alpha_1, ..., \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, ..., \alpha_k + x_k))$
$\mathcal{N}(\mu, \frac{1}{\theta})$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha + 0.5, \beta + \frac{(\mu-x)^2}{2})$

TABLE 3.1 – Lois a priori conjuguées usuelles

3.3 L'application de l'approche bayésienne aux modèles de mélange

L'utilisation des méthodes bayésiennes pour estimer les paramètres d'un mélange de distributions se justifie principalement par deux raisons :

Limites des méthodes classiques

Les méthodes telles que le maximum de vraisemblance se heurtent à des problèmes techniques lors de la modélisation des mélanges distributionnels. En fait, les probabilités de ces modèles sont généralement illimitées, de sorte que ces approches ne peuvent pas être appliquées directement.

L'approche bayésienne pour les modèles de mélange repose sur l'utilisation de méthodes d'inférence bayésienne pour estimer les paramètres de ces modèles. Cette approche permet de traiter des situations où les méthodes classiques comme le maximum de vraisemblance ne sont pas adaptées, notamment en présence d'observations aberrantes ou de distributions complexes.

Dans l'approche bayésienne, le paramètre inconnu θ est lui-même une variable aléatoire. Par conséquent, il est représenté par la probabilité π sur Θ, appelée probabilité a priori.

Selon une perspective bayésienne, la densité du modèle de mélange mentionné au premier chapitre devient :

$$g(x \mid \psi) = \sum_{k=1}^{K} p_k f(x \mid \theta_k), \tag{3.4}$$

où le paramètre ψ est défini comme $\psi = (\theta_1, \ldots, \theta_K, p_1, \ldots, p_K)$ et f est une fonction de densité et θ_i sont des paramètres, avec $p_1, ..., p_k$ représentant les coefficients ou proportions du modèle de mélange.

La fonction de vraisemblance, pour n observations $x_1, .., x_n$, associée au modèle (3.4) est :

$$\prod_{i=1}^{n} g(x_i \mid \psi) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f(x_i \mid \theta_k). \tag{3.5}$$

Dans le cadre du modèle de mélange, une distribution a priori est également spécifiée pour le paramètre ψ, notée $\pi(\psi)$. Il est important de souligner qu'il n'est généralement pas possible de choisir une distribution a priori impropre pour ψ, car cela pourrait conduire à une distribution a posteriori qui n'est pas une véritable mesure de probabilité. Cette contrainte empêche l'utilisation des distributions a priori non informatives usuelles, comme la loi de Jeffreys. Par conséquent, le choix de la distribution a priori revêt une importance capitale dans le contexte des modèles de mélange.

La loi a posteriori est définie par :

$$\pi(\psi \mid x_1, ..., x_n) = \frac{\pi(\psi) \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f(x_i \mid \theta_k)}{\int \pi(\psi) \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f(x_i \mid \theta_k) d\psi} \tag{3.6}$$

Cependant, le calcul de cette loi a posteriori peut être difficile ou impossible en pratique, notamment en raison de l'intégrale du dénominateur. Les approximations numériques peuvent être trop longues à calculer en raison du grand nombre de termes au numérateur et

au dénominateur. Les méthodes de Monte-Carlo par Chaînes de Markov (MCMC) ont été développées pour résoudre ces problèmes.

Exemple 3.3. *Un modèle de mélange gaussien :*

Supposons que nous ayons un échantillon de données $x_1, \ldots, x_n$ qui suit un mélange de distributions gaussiennes :

$$f(x|\Theta) = \sum_{k=1}^{K} p_k \mathcal{N}(x|\theta_k)$$

avec

$$\theta = (\theta_1, \ldots, \theta_K), \quad \theta_k = (\mu_k, \sigma_k^2) \in \Theta,$$
$$p = (p_1, \ldots, p_K) \quad et \quad \sum_{k=1}^{K} p_k = 1. \quad Où \quad \Theta = \{(p_k, \mu_k, \sigma_k), k = 1 \ldots, K\}$$

où p_k est le poids du $k^{ème}$ composante, μ_k est la moyenne et σ_k^2 est la variance de la $k^{ème}$ composante.

Pour ce modèle, la vraisemblance est :

$$L_n(\Theta|x) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \mathcal{N}(x_i|\mu_k, \sigma_k^2)$$

L'estimation des paramètres d'un modèle de mélange gaussien dans un cadre bayésien peut se faire en utilisant les lois a priori conjuguées. Cette méthode consiste à choisir des lois a priori qui, combinées aux données, donnent des lois a posteriori de même type. Pour ce modèle, les lois conjuguées usuelles sont :

- $p_k \sim Dirichlet(\alpha)$, α est un hyperparamètre.

- $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0^2)$, où μ_0 et σ_0^2 sont des hyperparamètres.

- $\sigma_k^2 \sim Inverse\text{-}Gamma(a, b)$, où a et b sont des hyperparamètres de la variance σ_k^2.

L'approche classique pour utiliser la loi conjuguée décrite ci-dessus consiste à utiliser la loi conjuguée (voir 3.3) pour les paramètres θ_i :

$$\pi(\theta_i \mid \nu_i, \lambda_i) \propto e^{\theta_i . \nu_i - \lambda_i \psi(\theta_i)};$$

où ν_i a la même dimension que θ_i. Ce type de distribution est particulièrement bien adapté aux structures exponentielles, y compris les modèles de mélange gaussiens.

La loi de Dirichlet $\mathcal{D}_k(\alpha_1, \ldots, \alpha_k)$ est définie par la densité contenant la fonction gamma Γ :

$$f(p|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)}_1^{\alpha_1 - 1} \ldots p_K^{\alpha_K - 1} \mathbb{1}_{\{\sum_{i=1}^K p_i = 1\}}.$$

Connaissant ces lois a priori, on peut calculer la loi a posteriori jointe des proportions p et des paramètres θ des composantes, conditionnellement aux données x :

$$\pi(p, \theta|x) \propto \mathcal{D}_k(\alpha_1, \ldots, \alpha_k) \prod_{k=1}^K \pi(\theta_k \mid \nu_k, \lambda_k) \prod_{i=1}^n \left(\sum_{k=1}^K p_k f(x_i|\theta_k) \right).$$

Cependant, cette loi a posteriori se décompose en K^n termes, rendant les calculs d'espérances a posteriori très coûteux pour de grands échantillons. Ce problème pratique pousse à utiliser d'autres méthodes d'estimation, comme l'algorithme EM ou l'apprentissage variationnel bayésien.

Pour plus de détails sur l'approche bayésienne d'un modèle de mélange, on peut voir par exemple Saint Pierre [2003].

Conclusion 3.4. *L'approche par les lois conjuguées permet une estimation bayésienne des paramètres de mélanges gaussiens, mais devient vite limitée en pratique. D'autres techniques sont alors nécessaires pour traiter efficacement ce type de modèles.*

Chapitre 4

Etude de simulation

4.1 Présentation du modèle de mélange gaussien

Dans cette section, nous présentons un exemple de simulation d'un mélange de deux lois gaussiennes bidimensionnelles, illustrant un processus de sélection aléatoire entre elles. En utilisant une loi de Bernoulli $\mathcal{B}(\lambda)$ pour choisir entre $\mathcal{N}_2(\mu_1, \Sigma_1)$ et $\mathcal{N}_2(\mu_2, \Sigma_2)$. Les fonctions $rbinom(n, size, prob)$ et $rnorm(n, mean, sd)$ du langage R sont employées pour générer des échantillons aléatoires issus des lois binomiale et gaussienne, respectivement. Par exemple, l'appel $rbinom(20, 8, 0.6)$ produira 20 observations d'une loi binomiale $\mathcal{B}(8, 0.6)$.

Nous générons ensuite un nuage de points, représentant un mélange de deux distributions gaussiennes bidimensionnelles.

$$\mathcal{N}_2\left(\begin{pmatrix}2\\1\end{pmatrix}, \frac{2}{3}\mathbf{I}_{\mathbb{R}^2}\right), \quad \mathcal{N}_2\left(\begin{pmatrix}8\\6\end{pmatrix}, \mathbf{I}_{\mathbb{R}^2}\right)$$

$$\text{avec} \quad \lambda = \frac{3}{5}$$

La figure ci-après illustre le nuage de points ainsi que l'histogramme correspondant à un échantillon de 100 observations issues d'un modèle de mélange de deux lois gaussiennes bidimensionnelles.

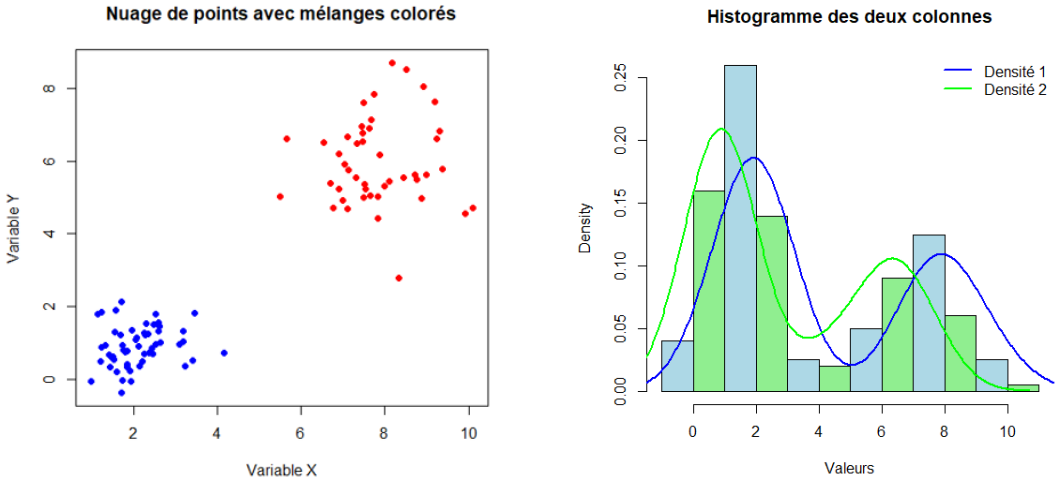

FIGURE 4.1 – Nuage de points et l'histogramme de mélange.

4.1.1 Application de l'algorithme EM pour l'estimation des paramètres du mélange gaussien unidimensionnel

L'algorithme EM est utilisé pour estimer les paramètres de deux mélanges gaussiens à partir de données artificielles. En utilisant une loi de Bernoulli $\mathcal{B}(1/2)$ pour sélectionner entre la première densité $\mathcal{N}_1(220, 3)$ et la deuxième densité $\mathcal{N}_2(250, 2)$, nous effectuons 100 itérations. Ce modèle est illustré à l'aide d'une représentation graphique, et les estimations des paramètres sont fournies avec l'erreur quadratique moyenne (RMSE) ainsi que l'écart-type (std).

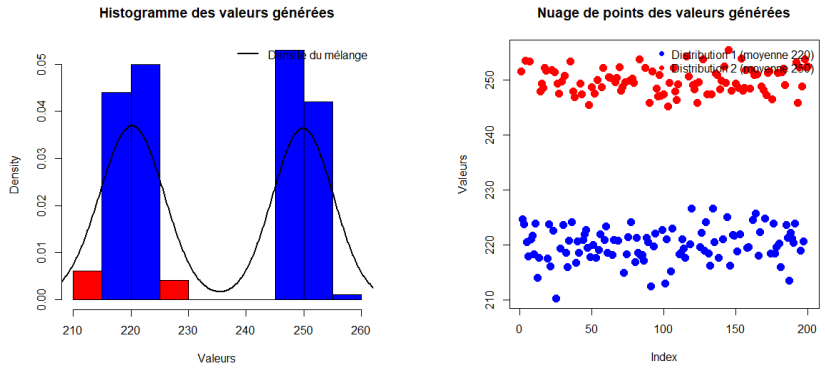

FIGURE 4.2 – Nuage de points et histogramme pour $n = 200$

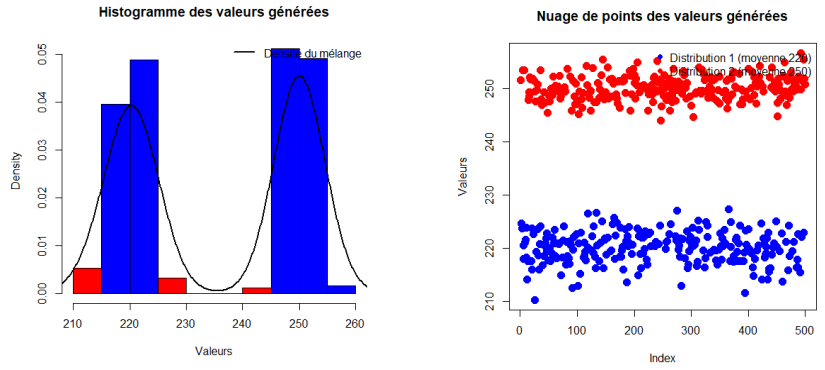

FIGURE 4.3 – Nuage de points et histogramme pour $n = 500$

Les résultats des estimations sont obtenus pour différentes tailles d'échantillons :

Paramètres	**Vraies valeurs**	$n = 30$	$n = 200$	$n = 500$
$\widehat{\mu_1}^{(\text{std})}_{(\text{rmse})}$	220	$220.7545^{(0.7850)}_{(0.7545)}$	$220.5213^{(0.3042)}_{(0.5213)}$	$219.9727^{(0.1876)}_{(0.0273)}$
$\widehat{\mu_2}^{(\text{std})}_{(\text{rmse})}$	250	$249.3674^{(0.7034)}_{(0.6326)}$	$250.0853^{(0.1642)}_{(0.0853)}$	$250.0662^{(0.1307)}_{(0.0662)}$
$\widehat{\sigma_1}^{(\text{std})}_{(\text{rmse})}$	3	$2.7098^{(0.5551)}_{(0.2902)}$	$2.8558^{(0.2151)}_{(0.1442)}$	$2.9840^{(0.1326)}_{(0.0160)}$
$\widehat{\sigma_2}^{(\text{std})}_{(\text{rmse})}$	2	$1.5425^{(0.4973)}_{(0.4575)}$	$2.2100^{(0.1161)}_{(0.2100)}$	$2.0331^{(0.0924)}_{(0.0331)}$
$\widehat{\lambda_1}^{(\text{std})}_{(\text{rmse})}$	0.5	$0.6333^{(0.0471)}_{(0.1333)}$	$0.5650^{(0.0500)}_{(0.0650)}$	$0.5120^{(0.0499)}_{(0.0120)}$
$\widehat{\lambda_2}^{(\text{std})}_{(\text{rmse})}$	0.5	$0.3667^{(0.0471)}_{(0.1333)}$	$0.4350^{(0.0500)}_{(0.0650)}$	$0.4980^{(0.0499)}_{(0.0120)}$

TABLE 4.1 – Résultats des estimations des paramètres $(\mu_1, \mu_2, \sigma_1, \sigma_2, \lambda_1, \lambda_2)$ d'un modèle de mélange de deux lois gaussiennes unidimensionnelles pour différentes tailles d'échantillons.

On remarque que la taille de l'échantillon influence les résultats de l'estimation. Plus la taille de l'échantillon est grande, plus les estimations des paramètres sont précises. En effet, le rmse et le std des estimations deviennent plus petits quand la taille de l'échantillon augmente.

Pour tester la normalité des observation, on a utilisé le Q-Q Plot et le Boxplot pour différents modèles de mélange de deux gaussiennes pour $n = 200$.

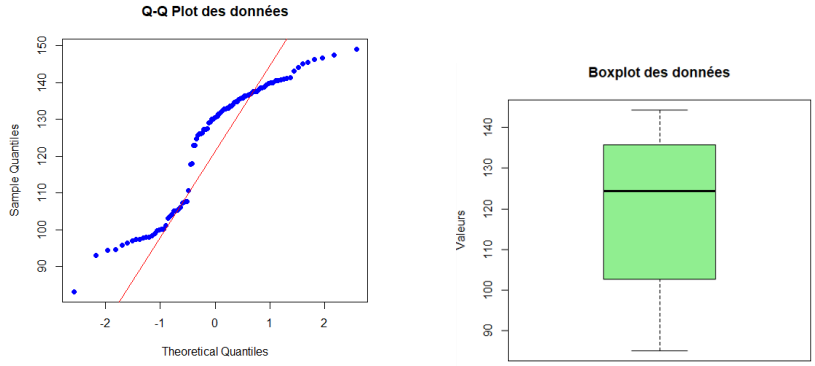

FIGURE 4.4 – Q-Q Plot et Boxplot d'un mélange de deux gaussiennes : "0.4 $\mathcal{N}_1(100, 6)+$ $0.6\mathcal{N}_2(133, 7)$"

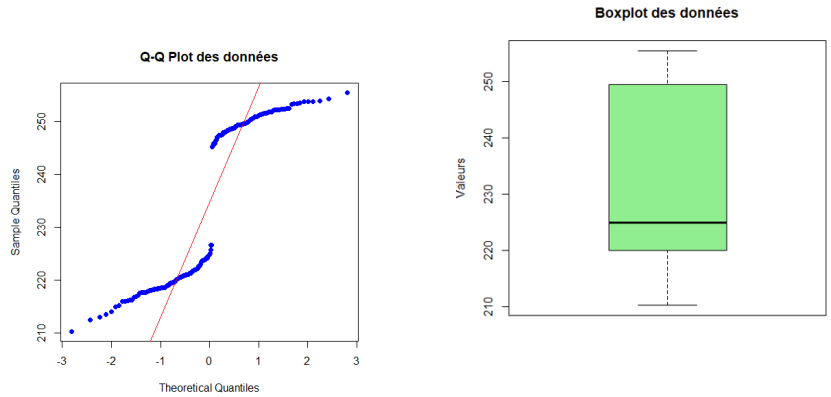

FIGURE 4.5 – Q-Q Plot et Boxplot d'un mélange de deux gaussiennes : "0.5 $\mathcal{N}_1(220, 3)+$ 0.5 $\mathcal{N}_2(250, 2)$"

On note que le modèle de mélange de gaussiennes ne présente pas de caractéristiques gaussiennes c'est à dire le mélange n'est pas gaussien.

Conclusion générale

Dans ce mémoire, nous avons examiné en profondeur la méthodologie classique pour estimer les paramètres d'un modèle de mélange.

Nous nous sommes concentrés sur les modèles de mélange paramétriques, en mettant particulièrement l'accent sur les mélanges gaussiens. Nous avons détaillé les méthodes classiques d'estimation des paramètres, en soulignant l'importance de l'algorithme EM. Par ailleurs, nous avons également exploré l'approche bayésienne des modèles de mélange. En intégrant des lois a priori sur les paramètres, cette approche offre une flexibilité accrue et permet d'incorporer des connaissances préalables dans le processus d'estimation.

Une étude de simulation a été présenté pour estimer les paramètres d'un modèle de mélange gaussien pour différentes tailles d'échantillons et de visualiser les données issues de ce modèle.

Plusieurs perspectives peuvent être explorées :

- Exploration d'autres algorithmes d'estimation en plus de EM, comme les méthodes Monte Carlo par chaînes de Markov (MCMC) qui permettent une estimation bayésienne plus flexible.
- Extension aux mélanges de distributions non gaussiennes (loi de Poisson, loi exponentielle, etc).

Résumé

الملخص

في هذا العمل، استخدامنا خوارزمية التوقع -التعظيم (EM) لتقدير معلمات نموذج خليط غاوسي من بيانات غير مكتملة. الهدف هو تحديد تأثير أحجام العينات المختلفة على جودة التقديرات.

Résumé

Dans ce travail, nous avons utilisé l'algorithme Expectation-Maximization (EM) pour estimer les paramètres d'un modèle de mélange gaussien à partir de données incomplètes. L'objectif est d'évaluer l'impact de différentes tailles d'échantillons sur la qualité des estimations.

Abstract

In this work, we used the Expectation-Maximization (EM) algorithm to estimate the parameters of a Gaussian mixture model from incomplete data. The goal is to evaluate the impact of different sample sizes on the quality of the estimates.

Mots clés : Modèle de mélange, Données incomplètes, Algorithme EM, Estimation des paramètres, Mélange de gaussiennes, Approche bayésienne.

Bibliographie

A. Gelman, J. B. Carlin, H. S. Stern, et D. B. Rubin. Analyse de données Bayésienne. Notes de cours. Chapman and Hall/CRC, 1995.

A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum de vraisemblance à partir de données incomplètes via l'algorithme EM. Journal of the Royal Statistical Society. Series B (Methodological), 39(1) :1–38, 1977.

B. P. Carlin et T. A. Louis. Méthodes Bayésiennes et empiriques pour l'analyse des données, vol. 7. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

C. Biernacki. Pourquoi les modèles de mélange pour la classification. CNRS Université de Lille 1, Villeneuve d'Ascq, France, biernack@math.univ-lille1.fr, 2009.

C. P. Robert. Méthodes de Monte Carlo par chaînes de Markov. Éditions Économica, p. 97, 1996.

C. Robert. Le choix bayésien : principes et pratique. Springer-Verlag, France, 2006.

D. Danho. Modèle de mélange et classification. Mémoire de Master, Université Paris-Dauphine, 2016.

D. Paul McNicholas. Clustering basé sur des modèles. Journal of Classification, 33(3) :331-373, 2016.

G. J. McLachlan et D. Peel. Modèles de mélange finis. Wiley Series in Probability and Statistics, Wiley-Interscience Publication, pp. 6, 29, 2000.

G. Saint Pierre. Identification du nombre de composants d'un mélange gaussien par chaînes de Markov à sauts réversibles dans le cas multivarié ou par maximum de vraisemblance dans le cas univarié. Thèse de Doctorat, Université Toulouse III, 2003.

J.J. Droesbeke, G. Saporta et C. Thomas-Agnan. Modèles à variables latentes et modèles de mélange. TECHNIP OPHRYS EDITIONS, 2013.

J. Kwon. Constantine Caramanis Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR 108 :1727-1736, 2020.

M. Haugh. The EM Algorithm. IEOR E4570 : Machine Learning for ORFE, pp. 1-7, 2015.

M. Collins. L'algorithme EM. www.cse.unr.edu/ bebis/CS679/Readings/EM Algorithm Review.pdf, septembre 1997.

R. C. P. Robert et G. Casella. Méthodes statistiques de Monte Carlo. Springer-Verlag, p. 31, 1999.

S. Morgenthaler. Génétique statistique. Collection « Statistique et probabilités appliquées », Springer, 2008.

S. Nefkha-Bahri. Modèle de mélange gaussien à effets superposés pour l'identification de sous-types de schizophrénie : Algorithme EM ; clustering ; algorithme EM, 2020.

T. Korbaa. A mathematical model for cyclic scheduling with work-in-progress minimization. 12th IFAC Symposium on Control, 2006.

V. Melnykov, R. Maitra, et al. Modèles de mélange finis et clustering basé sur des modèles. Statistics Surveys, pp. 80–116, 2010.