MODELE DE MELANGE ET CLASSIFICATION

Sous la direction de madame
ANGELINA ROCHE

Fait par DANHO DJROBIE

TABLE DES MATIERES

Introduction

La classification est un aspect important de l'analyse des données exploratoire et décisionnelle. Elle a pour objectif de déterminer si un ensemble d'objets est homogène et, lorsqu'il ne l'est pas, d'établir une partition de cette collection en sous-ensembles homogènes appelés **classes**: d'où le nom **classification.**

Néanmoins le critère d'homogénéité n'est pas toujours explicite ou unanime. D'où le problème pour établir des classes. Il est donc nécessaire de disposer d'outils s'adaptant à la nature des données à analyser.
Un outil populaire aujourd'hui dans la classification réside dans les modèles de mélanges.
Les modèles de mélanges, apparus dans les travaux de Pearson en 1984, sont utilisés avec succès dans bon nombre de disciplines comme l'astronomie, la biologie, la génétique, l'économie, les sciences de l'ingénieur, le marketing, la reconnaissance d'images...
L'on pourrait donc se demander : quel est le rôle joué par les modèles de mélanges dans la classification des données?
Notre problématique tend ainsi vers une approche probabiliste de classification où l'objet à classifier est l'échantillon d'un vecteur aléatoire.
Plus précisément : l'analyse de la densité d'une loi mélange.
Afin de répondre donc à cette problématique nous présenterons dans la section 1 les modèles de mélange.
Ensuite, dans la section 2 nous aborderons les modèles de mélanges paramétriques plus particulièrement l'estimation des paramètres à travers l'algorithme EM.
Enfin, dans la section 3 nous verrons une approche semi paramétrique des modèles de mélange ainsi qu'une approche bayésienne.

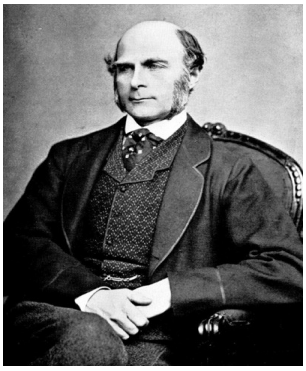
1-PRÉSENTATION GÉNÉRALE DES MODÈLES DE MÉLANGES

1.1-Contexte historique

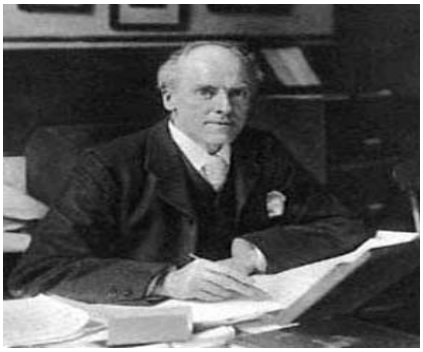
De nombreux scientifiques ont participé à l'émergence des statistiques et des probabilités. Parmi ces derniers, l'on compte Siméon Denis Poisson, Adolphe Quetelet, Francis Galton, Karl Pearson.

Siméon Denis Poisson a énormément contribué au développement des statistiques et des probabilités au cours du 19 ème siècle. Il a participé à la conception de nombreux résultats mathématiques : dans son travail sur les proportions de naissances des filles et des garçons, il établit la célèbre loi probabiliste qui porte son nom, à savoir la loi de Poisson. Il a également redémontré le théorème central limite.

Francis Galton s'est également intéressé aux mélanges de lois binomiales dans l'objectif de modéliser la transmission des caractères héréditaires sous l'influence de Darwin. En 1875, il écrit : «Pourquoi un mélange de séries radicalement différentes donnerait dans de nombreux cas des résultats apparemment identiques à ceux d'une simple série ?»



En ce qui concerne **Karl Pearson**, il met en évidence graphiquement en 1894 des modèles de mélange dans son article intitulé Contributions to the Mathematical Theory of Evolution.



1.2-Modèle de base

Le modèle de mélange fini de lois de probabilité consiste à supposer que les données proviennent d'une source contenant plusieurs sous-populations homogènes appelées composants.

La population totale est un mélange de ces sous-populations. Le modèle résultant est un modèle de mélange fini.

Soit $X=(X_1,\dots,X_n)$ un échantillon de variables aléatoires indépendantes identiquement distribués (iid) de loi mélange fini à K composants, de densité f dont

la forme générale est :

$$f(x)=\sum_{k=1}^{K}\Pi_k f_k(x) \quad (1.1)$$

avec : Π_k les proportions respectives des sous populations telle que $0<\Pi_k\leqslant 1 \; et \sum_{k=1}^{K}\Pi_k=1$

f_k la densité du kième composant (la paramétrisation des densités des composants dépend de la nature continue ou discrète des données observées).

Le modèle de mélanges est un modèle à données manquantes, en effet si on échantillonnait dans une population formée de K sous-populations, on devrait avoir les couples (X_i,Z_i)

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. En échantillonnant uniquement dans la sous population k et en supposant X discrète , on obtiendrait le modèle

$P(X=x\backslash Z=k)=f_k(x,\alpha_k)$ mais le paramètre α_k est en général inconnu et propre à la kième sous population.

De même, dans les modèles de mélanges les données manquantes sont $Z=(Z_1,\dots,Z_n)$ avec $Z_i=k$ si i provient du groupe k.

L'on observe donc que l'échantillon $(X_1, X_2, ..., X_n)$. Le mélange (1.1) peut être vu comme la loi marginale de la variable X pour le couple (X , Z). C'est donc un modèle à données manquantes.

1.3-Conditions d'identifiabilité des modèles de mélange

Il faut noter que pour pouvoir estimer des paramètres, il faut que les paramètres soient **identifiables** c'est à dire si $\Theta \neq \Theta' \; alors \; P_\Theta \neq P_{\Theta'}$.

En ce qui concerne les mélanges, les paramètres ne sont pas identifiables au sens classique.

Il faut donc des conditions d'identifiabilités spécifiques aux mélanges.

On pourra dire que H= $\{H(.)/H(x)=\sum_{k=1}^{K}\Pi_k F(x,\alpha_k), \Pi_k \geqslant 0, \sum_{k=1}^{K}\Pi_k = 1\}$ est identifiable si

$H=\sum_{k=1}^{K}\Pi_k F(.,\alpha_k) \equiv H'=\sum_{k=1}^{K}\Pi'_k F(.,\alpha'_k)$ implique $\sum_{k=1}^{K}\Pi_K \delta_{\alpha_k}=\sum_{k=1}^{K}\Pi'_K \delta_{\alpha'_k}$, autrement dit, il existe une permutation σ de { 1, … , K } telle que pour tout k, on a $\Pi_k=\Pi'_{\sigma_{(k)}}$ et $\alpha_k=\alpha'_{\sigma_{(k)}}$

avec $F(.,\alpha_k)$ la fonction de répartition du kième composant du mélange.

Il s'agit d'une notion d'identifiabilité à permutation des classes près.

Proposition

Les modèles de mélanges gaussiens univariés $\{H(.)/H(x)=\sum_{k=1}^{K}\Pi_k F(x,\alpha_k), \Pi_k \geqslant 0, \sum_{k=1}^{K}\Pi_k = 1\}$ sont identifiables.

Avec $F(.,\alpha_k)$ la fonction de répartition d'une gaussienne de densité

$f(x,\alpha_k)=\frac{1}{(2\pi\sigma_k^2)^{1/2}} e^{-(x-u_k)^2/2\sigma_k^2}$ et $\alpha_k=(\mu_k,\sigma_k)$

Preuve

La preuve s'appuie sur le théorème suivant (admis) :

Théorème (cf [1])

La famille H= $\{H(.)/H(x)=\sum_{k=1}^{K}\Pi_k F(x,\alpha_k), \Pi_k \geqslant 0, \sum_{k=1}^{K}\Pi_k = 1\}$ (avec $F(.,\alpha_k)$ la fonction de répartition à partir de laquelle est formée le mélange) est identifiable si et seulement si l'image de F par tout isomorphisme défini sur le sous espace vectoriel engendré par l'ensemble des fonctions de répartitions *F* est une famille libre dans l'espace d'arrivée.

Dans le cadre des mélanges gaussiens, l'isomorphisme qui nous intéresse est celui qui aux fonctions de répartitions associe les transformées de Laplace correspondantes.

Rappelons la formule de la transformée de Laplace :

$L_X(z)=E(e^{zX})$

La transformée de Laplace d'une gaussienne de densité $f(x,\alpha_k)=\frac{1}{(2\pi\sigma_k^2)^{1/2}}e^{-(x-u_k)^2/2\sigma_k^2}$ est :

$\exp(\mu_k z+\frac{1}{2}\sigma_k^2 z^2)$

On veut montrer que la famille des transformées de Laplace d'une gaussienne de densité $f(x,\alpha_k)=\frac{1}{(2\pi\sigma_k^2)^{1/2}}e^{-(x-u_k)^2/2\sigma_k^2}$ est libre.

Supposons donc que l'on ait : $\sum_{k=1}^{K}\lambda_k\exp(\mu_k z+\frac{1}{2}\sigma_k^2 z^2)=0$ avec λ_k k=1,...,K des réels non tous nuls.

Sans perte de généralité, on suppose que $\sigma_1^2<...<\sigma_K^2$.

On a : $\frac{1}{e^{\mu_K z+\frac{1}{2}\sigma_K^2 z^2}}\sum_{k=1}^{K}\lambda_k e^{\mu_k z+\frac{1}{2}\sigma_k^2 z^2}=0$ équivaut à

$\sum_{k=1}^{K}\lambda_k e^{(\mu_k-\mu_K)z+\frac{1}{2}(\sigma_k^2-\sigma_K^2)z^2}=0$ équivaut à

$\lambda_K+\sum_{k=1}^{K-1}\lambda_k e^{(\mu_k-\mu_K)z+\frac{1}{2}(\sigma_k^2-\sigma_K^2)z^2}=0$

$(\mu_k-\mu_K)z+\frac{1}{2}(\sigma_k^2-\sigma_K^2)z^2$ est un polynôme de degré 2 à coefficient dominant négatif qui tend vers 0 quand z tend vers $+\infty$.

Ainsi quand z tend vers $+\infty$, on obtient $\lambda_K=0$ et par récurrence descendante on a finalement $\lambda_1=...=\lambda_K=0$.

2)MODÈLES DE MÉLANGE PARAMÉTRIQUES

De façon générale, 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 l'espace des paramètres est de dimension finie.

2.1-L'approche du maximum de vraisemblance(ML) : l'algorithme EM

Comme son nom l'indique, l'approche ML consiste à maximiser la vraisemblance c'est à dire maximiser $L(\Theta, X)=\prod_{i=1}^{n}\sum_{k=1}^{K}\Pi_k f_k(X_i, \alpha_k)$ ou de façon équivalente à maximiser la log vraisemblance $l(\Theta, X)=\sum_{i=1}^{n}\log(\sum_{k=1}^{K}\Pi_k f_k(X_i, \alpha_k))$ afin d'estimer le paramètre inconnu

Avec $\Theta=(\Pi_1, \ldots, \Pi_K, \alpha_1, \ldots, \alpha_K)$ le paramètre inconnu du modèle paramétrique

$X=(X_1, \ldots, X_n)$ l'échantillon

$f_k(., \alpha_k)$ la densité du kième composant du modèle paramétrique : le paramètre $\alpha_k \in R^d$

Toutefois, ce problème de maximisation ne peut être résolu analytiquement en raison de données cachées. Il faut donc trouver les solutions à l'aide d'algorithmes itératifs. Parmi ces algorithmes, figure **l'algorithme EM**.

Cet algorithme est du à Dempster, Laird et Rubin (1977). Il vise à fournir un estimateur lorsqu'il est impossible de calculer la solution en raison de la présence de données cachées ou manquantes ou plutôt, lorsque la connaissance de ces données rendrait possible l'estimation des paramètres.

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

- la phase « Expectation », souvent désignée comme « l'étape E », procède à l'estimation des données inconnues, sachant les données observées et la valeur des paramètres déterminée à l'itération précédente;
- la phase « Maximisation », ou « étape M », procède donc à la maximisation de la vraisemblance, rendue désormais possible en utilisant l'estimation des données inconnues effectuée à l'étape précédente, et met à jour la valeur du ou des paramètre(s) pour la prochaine itération.

L'algorithme garantit que la vraisemblance augmente à chaque itération, ce qui conduit donc à des estimateurs de plus en plus corrects.

De façon plus formelle et plus généralement dans le cas discret,

- Si l'on dispose d'observations i.i.d. $X=(X_1, \ldots, X_n)$ de vraisemblance notée $P(X \setminus \Theta)$ dont la maximisation est impossible.

- L'on considère des données cachées $Z=(Z_1,\dots,Z_n)$ dont la connaissance rendrait possible la maximisation de la « vraisemblance des données complètes », l $P(X,Z\backslash\Theta)$;
- Etant donné que ces données Z sont inconnues, l'on **estime la vraisemblance des données complètes** en prenant en compte toutes les informations connues : pour cela on choisit comme estimateur

  $E_{Z\backslash X,\Theta_m}[\log P(X,Z\backslash\Theta)]$ (c'est l'« étape E » de l'algorithme) ;
- L'on maximise enfin cette **vraisemblance estimée** pour déterminer la nouvelle valeur du paramètre (« étape M » de l'algorithme).

  Ainsi, le passage de l'itération m à l'itération m +1 de l'algorithme consiste à déterminer :

$$\Theta_{m+1}=argmax\, E_{Z\backslash X,\Theta_m}[\log P(X,Z\backslash\Theta)]$$

2.2-Application de l'algorithme EM aux mélanges paramétriques: cas des mélanges gaussiens

Le modèle de mélange gaussien est une combinaison linéaire de plusieurs composantes gaussiennes. Il est particulièrement utilisé dans le cas où les données en études ne peuvent pas être modélisées par une simple gaussienne. En d'autres termes, si la structure de données est composée naturellement de plusieurs groupes, il faut les représenter par un modèle de mélange gaussien plutôt qu'une simple distribution gaussienne.

Ils sont considérés comme un outil important, pour modéliser et traiter les données multimédia (images, audio, vidéo). Les modèles de mélange gaussien sont utilisés aussi dans le domaine de reconnaissance audio.
Ces mélanges qui supposent les populations conditionnelles distribuées selon une loi normale, suscitent un intérêt important en raison :

(i) de leur flexibilité
(ii) de leur faculté à approcher une grande variété de densités : on peut toujours modéliser des données continues par un mélange gaussien.
(iii) de leur usage mathématiquement simple et de la généralité de la loi normale qu'atteste le théorème central limite.

Prenons par exemple un modèle de mélange de deux lois gaussiennes.

Soit $X=(X_1,\dots,X_n)$ un échantillon i.i.d. d'observations issues d'un mélange de deux gaussiennes bidimensionnelles, et soit $Z=(Z_1,\dots,Z_n)$ la donnée cachée où Z_i détermine la distribution dont est issue X_i :

$$L(X_i\backslash Z_i=1)=N_2(\mu_1,\Sigma_1)$$
$$L(X_i\backslash Z_i=2)=N_2(\mu_2,\Sigma_2)$$

avec $P(Z_i=1)=\Pi_1\, et\, P(Z_i=2)=\Pi_2=1-\Pi_1$

Ne connaissant que X on cherche à estimer les 5 paramètres inconnus $\Theta=(\Pi_1,\mu_1,\Sigma_1,\mu_2,\Sigma_2)$

La vraisemblance des données complètes est :

$$L(X, Z \setminus \Theta) = \prod_{i=1}^{n} \sum_{j=1}^{2} 1_{Z_i = j} \Pi_j f_j(X_i)$$

où $f_j : R^2 \to R$ est telle que $f_j(x) = \frac{1}{2\pi \, det(\Sigma_j)^{1/2}} \exp(-1/2 (x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j))$ est une densité gaussienne bidimensionnelle de paramètres μ_j et Σ_j . La log-vraisemblance des données complètes obtenue est :

$$\log(L(X, Z \setminus \Theta)) = \sum_{i=1}^{n} [\sum_{j=1}^{2} 1_{Z_i = j} (\log(\Pi_j) - \log(2\pi) - 1/2 \log(det(\Sigma_j)) - 1/2 (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j))]$$

à chaque itération, l'étape E nécessite de définir la distribution de Z_j connaissant X_j et Θ_m . On définit :

$$\tilde{p_{i,j}} = P(Z_i = j \setminus X_i = x_i, \Theta_m) = \frac{\Pi_j f_j(x_i)}{\Pi_1 f_1(X_i) + \Pi_2 f_2(X_i)}$$

la probabilité pour que le point X_i soit issu de la distribution $f_j \equiv N(\mu_j, \Sigma_j)$ connaissant Θ_m . Alors, on a :

$$E_{Z \setminus X, \Theta}[\log(L(X, Z \setminus \Theta))] = \sum_{i=1}^{n} \sum_{j=1}^{2} \tilde{p_{i,j}} (\log(\Pi_j) - \log(2\pi) - 1/2 \log(det(\Sigma_j)) - 1/2 (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j))$$

La maximisation en Θ de cette expression conduit aux estimateurs suivants :

$$\Pi_j^{(m+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{i,j}}}{n}$$

$$\mu_j^{(m+1)} = \frac{\sum_{i=1}^{n} \tilde{p_{i,j}} x_i}{\sum_{i=1}^{n} \tilde{p_{i,j}}}$$

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

A la convergence de l'algorithme, on peut déduire une partition en rangeant chaque individu dans la classe maximisant la probabilité à posteriori $\tilde{p_{i,j}}$: il s'agit de la méthode du **MAP ou maximum à posteriori**

3-MODÈLE DE MÉLANGES SEMI-PARAMÉTRIQUES ET APPROCHE BAYESIENNE

3.1-Modèle de mélanges semi-paramétriques : cadre de travail et identifiabilité

On se restreint à un modèle de mélange à deux composants identifiable défini par :

(2) $G(x)=\lambda F(x-\mu_1)+(1-\lambda)F(x-\mu_2)$, x dans R.

avec comme paramètres inconnus la fonction de répartition F d'une loi symétrique, 2 paramètres de translations réels μ_1 et μ_2 et les proportions du mélange λ.

Ce modèle a été étudié par Hunter, Wang et Hettmansperger. Il s'agit d'un modèle semi-paramétrique c'est à dire un modèle dans lequel les paramètres inconnus peuvent être séparés en une partie fonctionnelle F et une partie euclidienne λ,μ_1,μ_2 .

Si la fonction de répartition F admet une densité f, alors la loi mélange admet une densité g définie par :

(3) $g(x)=\lambda f(x-\mu_1)+(1-\lambda)f(x-\mu_2)$, x dans R,

où θ=(λ,μ_1,μ_2) appartient à Θ=[0,1/2)×(R2\Δ) and Δ={(x,x);x dans R}.

Le modèle (2) est **identifiable** si :

(4) $\lambda F(x-\mu_1)+(1-\lambda)F(x-\mu_2)=\lambda' F'(x-\mu_1')+(1-\lambda')F'(x-\mu_2')$ pour tout x dans R implique $\lambda=\lambda',\mu_1=\mu_1',\mu_2=\mu_2'$ ou $\lambda=(1-\lambda'),\mu_1=\mu_2',\mu_2=\mu_1'$

pour 2 différents quadruplets (θ,F) et (θ′,F′) appartenant à $\Theta\times F$, où θ′ = (λ',μ_1',μ_2') et F est l'ensemble des fonctions de répartitions de lois symétriques.

Il suffit de considérer λ appartenant à [0, 1/2) car le modèle est alors invariant par permutation de (λ,μ_1) et ($(1-\lambda),\mu_2$).

Théorème

Si (λ,μ_1,μ_2,F) et $(\lambda',\mu_1',\mu_2',F')$ sont deux paramètres appartenant à [0,1/2) × (R^2 \Δ) × F satisfaisant (4), alors λ = λ′, $\mu_2=\mu_2'$, $\mu_1=\mu_1'$ si λ>0 et F = F′.

Preuve

Etape 1

Soit $(\sin(\alpha_1 t),\dots,\sin(\alpha_p t))$ une famille de p fonctions définies sur R.

Cette famille est libre si et seulement si :

(5) $\alpha_i \neq 0$ pour $1\leq i\leq p$ et $|\alpha_i|\neq|\alpha_j|$ pour $1\leq i<j\leq p$.

En effet, soit $\beta_1,\dots,\beta_p$. On a $\sum_{i=1}^{p}\beta_i\sin(\alpha_i t)=0$ \$ t Ø R.

En dérivant cette expression par rapport à t à l'ordre 1,3,...,2p−1,on obtient en t=0 le système d'équations linéaires : $\sum_{i=1}^{p}\beta_i\alpha_i^{2j+1}=0$ pour $0\leq j\leq p-1$. Matriciellement , on a $(AV)^T=0$ avec

$$A=\begin{pmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_p\\ \alpha_1^3 & \alpha_2^3 & \dots & \alpha_p^3\\ \dots & \dots & \dots & \dots\\ \alpha_1^{2p-1} & \alpha_2^{2p-1} & \dots & \alpha_p^{2p-1}\end{pmatrix} \text{ et } V=\begin{pmatrix}\beta_1\\ \beta_2\\ \dots\\ \beta_p\end{pmatrix}$$

La matrice A est une matrice de type Vandermonde . Son déterminant est donc différent de 0 si et seulement si (5) est satisfait.

On vient donc de montrer que la famille $(\sin(\alpha_1 t),\dots,\sin(\alpha_p t))$ est libre si (5) est satisfait.

Etape 2

Soient Φ et Φ′ les fonctions caractéristiques de F et F′ respectivement.

En calculant la fonction caractéristique des 2 cotés de (4), on obtient , pour tout tØR,

$$(\lambda\exp(it\mu_1)+(1-\lambda)\exp(it\mu_2))\Phi(t)=(\lambda'\exp(it\mu_1')+(1-\lambda')\exp(it\mu_2'))\Phi'(t) \quad (6)$$

Puisque F et F′ sont des fonctions de répartitions de lois symétriques, leurs fonctions caractéristiques sont des fonctions continues égales à 1 en t = 0 . Ainsi avec (6) , on obtient que

$$(\lambda\exp(it\mu_1)+(1-\lambda)\exp(it\mu_2))(\lambda'\exp(-it\mu_1')+(1-\lambda')\exp(-it\mu_2')) =$$
$$\lambda'^2+(1-\lambda')^2+2\cos(\mu_1'-\mu_2') \quad . (7)$$

La partie imaginaire de (7) est donc égale à 0 dans un voisinage de 0.

On a ainsi :

(8)
$$\lambda\lambda'\sin((\mu_1-\mu_1')t)+\lambda(1-\lambda')\sin((\mu_1-\mu_2')t)+(1-\lambda)\lambda'\sin((\mu_2-\mu_1')t)+(1-\lambda)(1-\lambda')\sin((\mu_2-\mu_2')t)=0$$
sur la ligne des réels par analyticité des fonctions sinus .

Cas 1: λ = 0.

(8) devient :

(9) $\lambda'\sin((\mu_2-\mu_1')t)+(1-\lambda')\sin((\mu_2-\mu_2')t)=0$.

si λ′ >0, alors on a 1−λ′ >λ′ >0, et par l'étape 1 on considère ces différents sous cas :

• si $\mu_2=\mu_2'$ ou $\mu_2=\mu_1'$, avec (9) on a $\mu_2'=\mu_1'$, (pas admissible).

• si $|\mu_2'-\mu_1'|=|\mu_2'-\mu_2'|$, avec (9) on a λ′ +(1−λ′)=0 (impossible)

ou λ′ − (1 − λ′) = 0 (pas admissible).

Ainsi λ′ =λ=0 et donc par (8) $\mu_2=\mu_2'$.

Cas 2: λ > 0.

par le cas 1, on a aussi λ′ > 0.

Ainsi, il reste à montrer que

si $\mu_1\neq\mu_2,\mu_1'\neq\mu_2'$, $(\lambda,\lambda')\in(0,1/2)^2$ et que pour tout tØR,

$$\lambda\lambda'\sin((\mu_1-\mu_1')t)+\lambda(1-\lambda')\sin((\mu_1-\mu_2')t)+(1-\lambda)\lambda'\sin((\mu_2-\mu_1')t)+(1-\lambda)(1-\lambda')\sin((\mu_2-\mu_2')t)=0$$

(10) on a $(\lambda,\mu_1,\mu_2)=(\lambda',\mu_1',\mu_2')$.

Si on note β_1 =λλ′, β_2 =λ(1−λ′), β_3 = λ′(1 − λ) et β_4 = (1 − λ)(1 − λ′), alors (10) est équivalent à

(11) β_1 sin(αt) + β_2 sin((α′ − η)t) + β_3 sin((α + η)t) + β_4 sin(α′ t) = 0 pour tout t dans R,

avec $\alpha=\mu_1-\mu_1'$, $\alpha'=\mu_2-\mu_2'$ et $\eta=\mu_2-\mu_1$.

Si α=α′ =0, alors (11) devient $\sin(\eta t)(\beta_3-\beta_2)=0$. On a donc λ=λ′.

Montrons que $(\alpha,\alpha')\neq(0,0)$ n'est pas admissible .

Considérons le cas $\alpha=0$ et $\alpha'\neq0$ .Le cas $\alpha\neq0$ et α′ = 0 se fait de manière symétrique. Ainsi si on suppose que $\alpha=0$ et $\alpha'\neq0$, (11) devient

$$\beta_2\sin((\alpha'-\eta)t)+\beta_3\sin(\eta t)+\beta_4\sin(\alpha' t)=0 \quad \text{pour tout t dans R. (12)}$$

Puisque α′ et η sont non nuls , par l'étape 1, on considère les sous cas suivants :

• α′ =η: ainsi, (β3 +β4)sin(ηt)=0 for all tØR. ainsi β3 +β4 =0, ce qui donne λ′(1 − λ) + (1 − λ)(1 − λ′)=0 ou encore (1 − λ)=0 impossible.

• |α′ − η| = |η|: comme α′ et η sont non nuls on a donc α′ − η = η ou α′ − η = -η (impossible)

ainsi, α′ = 2η et (12) devient (β2 +β3)sin(ηt)+β4sin(2ηt)=0 pour tout t dans R,

qui, encore par l'étape 1, ne peut pas être satisfait pour tout t dans R.

• Les cas |α′ −η|=|α′| et|η|=|α′| nous donnent α′ =η/2 et η= −α′,

ainsi, comme dans le cas précédent, les équations résultantes ne peuvent être satisfaites pour tout t dans R.

on a donc montré que $(\lambda,\mu_1,\mu_2)=(\lambda',\mu_1',\mu_2')$

Etape 3

On veut montrer que F = F′ . Ce qui revient à montrer que Φ = Φ′.

Puisque λØ[0,1/2) et $(\lambda,\mu_1,\mu_2)=(\lambda',\mu_1',\mu_2')$, avec(6) il suffit de montrer que : λexp(itμ1)+(1−λ)exp(itμ2) est non nul.

On a:

$|\lambda\exp(it\mu 1)+(1-\lambda)\exp(it\mu 2)|\geq 1-2\lambda.$

Ainsi $\Phi = \Phi'$ et finalement $F = F'$

3.2-Estimation des paramètres : algorithme EM non paramétrique

Il faut noter que l'algorithme EM non paramétrique ne correspond pas exactement à l'algorithme EM. En effet, il n'a pas été prouvé que cet algorithme puisse maximiser une vraisemblance.

Comme dans le cas paramétrique soit $X=(X_1,\dots,X_n)$ un échantillon i.i.d. d'observations, $Z=(Z_1,\dots,Z_n)$ la donnée cachée où Z_i détermine la distribution dont est issue X_i :

Z_i =j indique que l'individu i provient de la jième composante du mélange. Les données complètes sont donc les (X_i , Z_i), 1 ≤ i ≤ n.

Soient les données initiales $\varphi^0=(\lambda^0, f^0)$ où λ^0 représente les proportions respectives des sous populations à l'état initial et f^0 la densité de la loi mélange à l'état initial. Pour t = 1, 2, . . ., l'algorithme suit 3 étapes dont les 2 premières sont semblables aux étapes de l'algorithme EM paramétrique:

-<u>l'étape E</u> : Consiste à calculer les probabilités "à posteriori"(conditionnellement aux données et φ^t) ,

$$p_{ij}^t=P_{\varphi_t}(Z_i=j\setminus x_i)=\frac{\lambda_j^t f_j^t(x_i)}{\sum_{j=1}^{K}\lambda_j^t f_j^t(x_i)}$$

pour i = 1,...,n and j = 1,...,K.

-<u>l'étape M</u>:

pour j = 1,...,K, on pose :

$$\lambda_j^{t+1}=\frac{\sum_{i=1}^{n} p_{ij}^t}{n}$$

-<u>Estimation de la densité non paramétrique</u>

C'est cette étape qui diffère de l'algorithme EM paramétrique : plutôt que d'estimer les paramètres de la loi , on estime la densité à l'aide de la méthode des noyaux. L'**estimation par noyau** (ou encore méthode **de Parzen-Rosenblatt**) est une méthode non-paramétrique d'estimation de la densité d'une variable aléatoire. Elle se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support. En ce sens, cette méthode généralise astucieusement la méthode d'estimation par un histogramme.

On obtient :

$$\mu_j^t = \frac{\sum_{i=1}^{n} p_{ij}^t x_i}{n \sum_{i=1}^{n} p_{ij}^t}$$

$$f^{t+1}(u) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij}^t K(u - x_i + \mu_j^t / h)}{n \sum_{i=1}^{n} p_{ij}^t}$$

où K(t)= $\frac{1}{\sqrt{2\pi}} \exp(\frac{-1}{2} t^2)$ est le noyau gaussien par exemple.

3.3-L'approche bayésienne

Jusqu'à présent , nous raisonnions avec une approche fréquentiste.

Toutefois l'on pourrait également voir les modèles de mélange d'un point de vue Bayésien.

Dans l'approche bayésienne, le paramètre inconnu θ est lui même une variable aléatoire comme les observations. Il est donc représenté par une probabilité π sur Θ appelée probabilité à priori.

Soit le modèle de mélange de densité $g(x) = \sum_{i=1}^{K} p_i f(x \backslash \theta_i)$,

avec $p_1, \ldots, p_K$ les poids du modèle.

Les paramètres du modèle sont $\{ p_1, \ldots, p_K, \theta_1, \ldots, \theta_K, z_1, \ldots, z_n \}$ avec z_i indicateur de la composante d'origine de x_i .

On décompose la loi a priori de la forme suivante :

π(p,θ,z) = $\pi(z|p, \theta_1, \ldots \theta_K) \pi(\theta_1, \ldots \theta_K, p) = \pi(z|p) \pi(\theta_1, \ldots, \theta_K, p)$

où π(z|p) Å $p_1 1_{z=1} + \ldots + p_K 1_{z=K}$

La loi de z sachant $p, \theta_1, \ldots, \theta_K$ est indépendante de $\theta_1, \ldots, \theta_K$. $p, \theta_1, \ldots, \theta_K$ sont indépendants.

On estime à partir de la loi a posteriori de z_i la composante d'origine de l'observation x_i . Le critère de classification est le même que celui de l'approche fréquentiste :

On décide que l'observation x_i est issue de $f_{J(i)}$ où :

J(i) = $argmax_{j=1,\dots,K} P(z_i = j | x_1, \dots, x_n)$

Cas particulier : population à deux composantes

il s'agit d'un problème de test d'hypothèses tel que

hypothèse nulle : x_i est issue de la première composante

hypothèse alternative : x_i est issue de la deuxième composante

Il suffit de calculer $P(z_i = 1 | x_1, \dots, x_n)$. Si $P(z_i = 1 | x_1, \dots, x_n)$ > 1/2 alors on décide que la composante x_i est issue de la première composante.

Il est fréquent que la structure de modélisation d'où sont issues les données x soit elle-même incertaine. Par exemple, on ignore le nombre K de composantes du modèle et ce nombre doit donc être lui aussi estimé.

Dans cette logique, l'approche bayésienne peut être utile dans le choix du nombre de classes grâce au critère BIC.

CONCLUSION

Grâce à leur flexibilité et à différentes conditions d'identifiabilité, les modèles de mélange jouent un rôle considérable au niveau de la classification des données. En ce sens où, grâce à leur approche probabiliste, ils permettent de résoudre les problèmes de partition qui étaient plus complexes avec l'approche ancienne que constituait l'approche géométrique.

Les modèles de mélange à la fois paramétriques et semi paramétriques, grâce à l'approche EM et à l'approche du maximum à posteriori, constituent donc un outil de classification très prisé dans l'analyse de grandes données.

C'est notamment le cas des mélanges gaussiens qui sont utilisés dans le domaine de reconnaissance audio.

Néanmoins des recherches sont encore en cours afin d'améliorer les résultats d'identifiabilité particulièrement dans les modèles semi-paramétriques.

BIBLIOGRAPHIE

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

[2] Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. Journal of Computational and Graphical Statistics, 18, 505–526.

[3] Bordes, L., Mottelet S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. Annals of Statistics, 34, 1204–1232.