Université de Montréal

UNE APPROCHE BAYÉSIENNE DE LA CLASSIFICATION HIÉRARCHIQUE

par

Aurélie Labbe

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

En vue de l'obtention du grade de

Maître ès sciences (M. Sc.)
en mathématiques

février 2000

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

UNE APPROCHE BAYÉSIENNE DE LA CLASSIFICATION HIÉRARCHIQUE

présenté par

Aurélie Labbe

a été évalué par un jury composé des personnes suivantes :

Robert Cléroux
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Yves Lepage
(membre du jury)

Mémoire accepté le :

Le 11 mai 2000

SOMMAIRE

Dans le cadre de notre étude, nous développons une nouvelle méthode de classification hiérarchique bayésienne. Cette approche est itérative et consiste en une série de partitions ou de regroupements des données selon un critère d'homogénéité que nous introduisons et qui se compare avantageusement à beaucoup de méthodes usuelles classiques.

La définition d'homogénéité des données est basée sur la construction d'un test d'hypothèses vérifiant l'égalité des deux premiers moments des variables. Le modèle que nous proposons *a priori* est un modèle hiérarchique, dont les paramètres de nuisances sont estimés dans un cadre bayésien empirique. Le test que nous utilisons est alors construit à partir du calcul des distributions marginales sous chacune des quatre hypothèses d'égalité de moyennes et de variances. L'algorithme de classification nous amène alors à la formation de groupes, homogènes ou hétérogènes au niveau des moyennes et des variances. Cette approche, unidimensionnelle au départ, est par la suite généralisée au cas multidimensionnel. Une comparaison de la nouvelle approche avec les méthodes classiques les plus fréquemment utilisées est ensuite établie, après avoir auparavant détaillé chacune d'entre elles.

REMERCIEMENTS

Lorsque Jacques Cartier traversa l'Atlantique, son but était de rejoindre la Chine. Exactement 464 ans plus tard, je suivis sa voie dans le but de me " reposer " quelques mois. Décidement, il faut croire que le Québec détourne les visiteurs de leur route car je me retrouve aujourd'hui à exprimer ma reconnaissance à tous ceux qui m'ont aidé à rédiger ce mémoire, fruit de mes deux années de travail ici. Le premier d'entre eux est sans conteste mon directeur de recherche, Jean-François Angers, qui a su rendre si agréable ce projet de recherche. Il n'est d'ailleurs pas étranger à ma décision de poursuivre mes études au doctorat et je crois que l'expérience acquise en travaillant avec lui me sera très profitable dans le futur. On dit que la recherche, c'est à la fois inspiration et transpiration... Et bien merci pour toutes ces inspirations qui m'ont tant fait transpirer...

Je voudrais aussi remercier Robert Cléroux pour son aide précieuse et efficace lorsque les problèmes administratifs dus à mon statut d'étudiante étrangère me semblaient difficiles à surmonter. Merci également à André Montpetit pour son aide à la mise en page, à Janie et Isabelle d'avoir si gentiment répondu à toutes mes petites questions quotidiennes et à Miguel Chagnon, grâce à qui j'ai pu si agréablement concilier recherche et travail.

Enfin, merci à Marc pour son aide de tous les jours, pour ces petits plus que je n'énumèrerais pas, mais qui m'ont rendu mon travail tellement plus facile.

Finalement, Jacques Cartier a découvert le Canada...Moi, j'ai découvert des personnes accueillantes et chaleureuses avec qui j'ai eu un réel plaisir à travailler.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

Classifier est un acte que chacun d'entre nous effectue quotidiennement, à chaque instant, de façon inconsciente. Le langage en est un parfait exemple : lorsque nous voulons exprimer un mot ou une phrase, nous classons ensemble des lettres propres à former des syllabes, qui ensemble forment des mots. Dans le domaine de la recherche, la classification est aussi à la base de tous les progrès et ce, depuis des milliers d'années. En effet, un élément fondamental de la médecine, l'anatomie, n'est autre qu'une gigantesque classification du corps humain. Successivement, l'homme a appris à classifier les aliments, les plantes, les animaux...pour aboutir aujourd'hui à des systèmes de classes, de regroupement de plus en plus complexes et diversifiés. Quoi de plus naturel, donc, dans un processus d'analyse de données, que de vouloir obtenir un regroupement de celles-ci en classes (ou groupes) homogènes par rapport à un certain critère.

Il existe aujourd'hui un très grand nombre de méthodes scientifiques permettant de tels classements. Le but principal de notre mémoire est d'en proposer une nouvelle, s'appuyant sur des arguments bayésiens. Ces derniers arguments nous permettent, entre autres, d'utiliser l'information *a priori* que nous pouvons avoir sur les données et de la combiner à celle apportée par l'échantillon.

Nous pouvons remarquer, en étudiant les méthodes fréquentistes les plus couramment utilisées, que le critère principal de classification reste toujours relié aux valeurs que peuvent prendre les données. L'innovation principale que nous proposons avec le développement d'une méthode bayésienne est l'introduction d'un

critère relié à la variance des observations. Nous classons donc non seulement celles-ci en classes, à l'intérieur desquelles les moyennes des observations sont égales, mais aussi en classes, à l'intérieur desquelles les variances des observations sont égales. Ceci nous permet donc de combiner à la fois l'égalité de moyennes et de variances des observations à l'intérieur d'un même groupe.

Nous consacrons le premier chapitre de ce mémoire à une revue de littérature des principales méthodes fréquentistes de classifications usuelles. Celles-ci sont d'ailleurs toutes disponibles sur la plupart des logiciels statistiques. Toutes ces méthodes sont de type hiérarchique et agglomératif, c'est-à-dire que le processus de classification est algorithmique, de façon à ce qu'un regroupement de deux données ou de deux groupes de données ait lieu à chaque itération. De plus, le terme agglomératif signifie que l'algorithme débute avec un nombre de groupes égal au nombre d'observations (chaque groupe contient donc une unique observation), pour aboutir à une seule classe, contenant toutes les observations. Le nombre de groupes est alors diminué de un à chaque itération de l'algorithme. Après avoir brièvement introduit les mesures de similarité et de dissimilarité entre les variables utilisées en classification, nous détaillons huit différentes méthodes. Pour la plupart, la différence entre elles n'est qu'une question de choix du type de distance utilisé. Ainsi, la méthode du plus proche voisin choisit de regrouper à chaque étape les deux observations (ou les deux classes d'observations) les plus proches en terme de distance. Au contraire, la méthode du voisin le plus éloigné regroupe les observations les plus éloignées. La méthode de la moyenne, quant à elle, se rapproche de la méthode du plus proche voisin au sens où elle regroupe les groupes dont la distance est minimum, mais la notion de distance inter-groupe est plus raffinée et se base sur la moyenne des distances entre tous les couples possibles d'observations appartenant à ces groupes. Ceci diffère encore

de la méthode centroïde qui, dans la même situation, choisit plutôt de prendre la distance entre les centroïdes respectifs des deux classes que l'on veut regrouper. La méthode de la médiane, basée sur la méthode centroïde, diffère de celle-ci par le simple fait qu'elle ne tient pas compte de la taille des groupes qu'elle veut rassembler. La méthode de Ward est quant à elle plus particulière : elle associe à chaque classification un coût que l'on veut, bien sûr, minimiser. Ceci reste dans l'optique des méthodes d'optimisation, qui minimisent certains critères relatifs à la variance à l'intérieur de chaque classe. Enfin, la dernière méthode fréquentiste que nous analysons est celle du maximum de vraisemblance, qui se base sur des modèles mixtes.

Comme nous pouvons donc le remarquer, les méthodes fréquentistes sont nombreuses et très diversifiées. Il n'en est pas de même malheureusement pour les méthodes bayésiennes, que nous étudions dans le deuxième chapitre, qui sont souvent développées dans un contexte très particulier ou dans le cas d'une étude précise. Pourtant, beaucoup de ces méthodes sont basées, ou font référence à celle développée par Binder en 1978 et celle proposée par Symons en 1981. La première est tout simplement une justification bayésienne des critères utilisés par les méthodes d'optimisation fréquentistes et la seconde est une extension de la méthode du maximum de vraisemblance, donc basée elle aussi sur des modèles mixtes. Ces méthodes sont beaucoup plus complexes et intéressantes d'un point de vue statistique que beaucoup de méthodes fréquentistes.

La méthode que nous développons au troisième chapitre se détache des autres méthodes par le fait que la variance des observations devient un critère de classement au même titre que la moyenne. Cette méthode est basée sur des tests d'hypothèses que nous avons construits, sous un certain modèle *a priori* . Ceux-ci nous permettent de tester l'égalité des moyennes de deux observations (ou de

deux groupes d'observations), ou bien de tester à la fois l'égalité de moyenne et de variance. Ceci nous amène donc à la formation de quatre types de groupes : ceux dont la moyenne et la variance sont égales, ceux dont la moyenne est différente mais la variance égale, ceux qui, au contraire, ont la même moyenne mais une variance différente et enfin, ceux dont la moyenne et la variance sont différentes. Cette méthode a l'avantage, outre de raffiner la classification usuelle avec l'introduction du deuxième moment, de ne pas considérer le nombre de groupes comme une variable connue et fixe ce qui, comparativement aux autres méthodes étudiées, est une innovation. Toutes les méthodes que nous avons développées sont appliquées à un jeu de données bien connu en classification : les iris de Fisher. Nous établissons alors une comparaison des méthodes au dernier chapitre.

Chapitre 1

MÉTHODES CLASSIQUES

Dans ce premier chapitre, nous définissons tout d'abord la notion de mesure de similarité et de dissimilarité entre 2 individus. Nous exposons les différentes mesures usuelles en classification, ainsi que leurs particularités. Par la suite, nous étudions les méthodes fréquentistes les plus courantes : les méthodes agglomératives et les méthodes d'optimisation. Ce chapitre est principalement tiré du sixième chapitre du livre de Lorr (1983), ainsi que des chapitres 3, 4 et 5 de Everitt (1993).

1.1. Les mesures de dissimilarité

Toute classification en tant que telle ne débute pas avec la donnée d'une série de variables. Une étape préliminaire essentielle consiste en fait à construire une matrice dont les éléments, appelés indices de proximité, indiquent la similarité ou dissimilarité de toutes les paires d'individus que nous cherchons à regrouper. Il existe plusieurs façons de construire un tel indice. Nous détaillons dans cette section plusieurs distances utilisées couramment en classification en considérant la donnée de n variables à p dimensions $X_1, X_2, ..., X_n$.

Définition 1.1.1. *Un coefficient de dissimilarité entre 2 variables X_i et X_j, noté d_{ij}, est une fonction non négative de leurs valeurs observées,*

$$d_{ij} = f(X_i, X_j).$$

Ce coefficient satisfait les 4 axiomes suivants :

- $\forall i, j \in \{1, ..., n\}, \quad d_{ij} \geq 0,$
- $\forall i \in \{1, ..., n\}, \quad d_{ii} = 0,$
- $\forall i, j \in \{1, ..., n\}, \quad d_{ij} = d_{ji},$
- $\forall i, j, k \in \{1, ..., n\}, \quad d_{ij} + d_{ik} \geq d_{kj}.$

Nous constatons aisément que cette notion de dissimilarité est équivalente à la notion de distance dans un espace métrique. Mais cette appellation particulière est justifiée par la démarche suivante.

Définition 1.1.2. *Un coefficient de similarité entre 2 variables X_i et X_j est une fonction de leurs valeurs observées indiquant leur degré de relation. Nous le notons*

$$s_{ij} = f(X_i, X_j).$$

Ce coefficient admet la propriété de symétrie suivante :

$$s_{ij} = s_{ji} \quad \forall i, j \in \{1, ..., n\}.$$

Il est courant en classification de construire un tel coefficient se comportant comme un coefficient de corrélation. Nous avons alors :

$$-1 \leq s_{ij} \leq 1 \quad \forall i, j \in \{1, ..., n\}.$$

L'indice de dissimilarité est alors construit de façon complémentaire :

$$d_{ij} = 1 - s_{ij}.$$

Il s'en suit alors que $s_{ii} = 1 \Leftrightarrow d_{ii} = 0, \ \forall i \in \{1, ..., n\}$. Les 4 axiomes de distance sont alors satisfaits par d_{ij}.

TABLEAU 1.1.1. *Information apportée par 2 variables dichotomiques*

$X_2 \backslash X_1$	0	1	Total
0	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+c$	p

1.1.1. Mesure de similarité pour les variables dichotomiques

Étudions maintenant la construction d'indices de similarité dans le cas de variables dichotomiques, définies comme étant des variables à 2 modalités uniques, que nous notons 0 et 1. Soient 2 variables dichotomiques X_1 et X_2 à p dimensions. L'information apportée par ces 2 variables peut être résumée dans un tableau 2×2 (voir tableau 1.1.1).

Un tel tableau est interprété de la façon suivante :

$$\begin{cases} a = \#\{i \text{ tels que } X_{1i} = X_{2i} = 0\}, \\ b = \#\{i \text{ tels que } X_{1i} = 1 \text{ et } X_{2i} = 0\}, \\ c = \#\{i \text{ tels que } X_{1i} = 0 \text{ et } X_{2i} = 1\}, \\ d = \#\{i \text{ tels que } X_{1i} = X_{2i} = 1\}, \end{cases}$$

où $\#A$ représente la cardinalité de l'ensemble A. Nous avons donc ici $p = a+b+c+d$.

Plusieurs coefficients de dissimilarité ont été proposés, combinant les quantités a, b, c et d :

i) $d_{12} = \dfrac{a+d}{p}$ " Coefficient de concordance simple ",

ii) $d_{12} = \dfrac{d}{b+c+d}$ " Coefficient de Jacquard ",

iii) $d_{12} = \dfrac{2d}{2d+b+c}$,

iv) $d_{12} = \dfrac{2(a+d)}{2(a+d)+b+c}$,

v) $d_{12} = \dfrac{d}{d+2(b+c)}$.

Ces coefficients s'appliquent dans le cas où l'une des deux composantes est une traduction numérique d'une variable qualititive représentant une négation. Par exemple, a représente la fréquence calculée lorsque les deux variables étudiées admettent " non " comme réponse. En ce qui a trait au coefficient de Jacquard, cette dernière composante est ignorée. Nous pouvons aussi dans certains cas " favoriser " le cas des composantes identiques en appliquant un poids double à ces valeurs : les trois mesures iii) et iv) en sont une illustration. La mesure v) favorise plutôt les composantes non identiques.

1.1.2. Mesure de dissimilarité pour les variables aléatoires mixtes

Dans certains cas, il peut être judicieux de transformer une variable quantitative $X \in \mathbb{R}^p$ en une variable binaire $Y \in \{0,1\}^p$ nous ramène alors au cas précédent. Soit $x_0 \in \mathbb{R}^p$, une constante fixée. Un tel type de transformation serait par exemple

$$Y_i = \begin{cases} 0 & \text{si } X_i \leq x_{0,i}, \\ 1 & \text{si } X_i > x_{0,i}, \end{cases}$$

où la transformation s'applique composante par composante.

Posons maintenant $X_1,...,X_n$ n variables telles que $X_i \in \mathbb{R}^p \ \ \forall i \in \{1,...,n\}$.

Supposons aussi que les p variables identifiant chaque X sont de types différents (dichotomiques ou non). Gower (1967) propose alors une mesure de similarité particulièrement utile :

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}},$$

où s_{ijk} est la mesure de similarité entre les individus i et j pour la variable k ($k = 1, ..., p$), et où w_{ijk} est un poids associé à la variable k (souvent égal à 0 ou 1) permettant de ne pas considérer une comparaison qui serait non pertinente ($w_{ijk} = 0$) pour une variable k particulière.

1.1.3. Mesure de dissimilarité pour des variables aléatoires quelconques

Plusieurs mesures de dissimilarité (distance) ont été proposées. La mesure la plus couramment utilisée est la distance euclidienne. Les mesures suivantes représentent les distances usuelles utilisées en classification :

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \text{ “ Distance euclidienne ”},$$

$$d_{ij} = \sum_{k=1}^{p} |X_{ik} - X_{jk}| \text{ “ Distance en norme } L_1 \text{ ” },$$

$$d_{ij} = 1 - \frac{\sum_{k=1}^{p} (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\left(\sum_{k=1}^{p} (X_{ik} - \bar{X}_i)^2 \right)^{1/2} \left(\sum_{k=1}^{p} (X_{jk} - \bar{X}_j)^2 \right)^{1/2}}$$

“ Coefficient de corrélation ” ,

$$d_{ij} = \frac{\sum_{k=1}^{p} X_{ik}X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2 \sum_{k=1}^{p} X_{jk}^2}} \text{ “ Séparation angulaire ”,}$$

où $\bar{X}_i$ représente la moyenne de toutes les composantes de l'individu i. De la même façon que nous avons défini les mesures de dissimilarité entre 2 variables, nous pouvons définir les mesures de dissimilarité entre 2 groupes A et B.
Dans le cas de la distance euclidienne, nous avons alors par exemple :

$$d_{AB} = \sqrt{\sum_{k=1}^{p} (\bar{X}_{Ak} - \bar{X}_{Bk})^2},$$

où $\bar{X}_{Ak}$ est la moyenne du groupe A sur la variable k ($k = 1, ..., p$). Beaucoup d'autres mesures tenant compte notamment de la corrélation entre les variables ont été proposées (voir Everitt (1993), p.49).

1.2. MÉTHODES AGGLOMÉRATIVES

Les méthodes agglomératives hiérarchiques sont très populaires. Il existe un nombre d'algorithmes considérable, mais ils peuvent être classés en trois grandes catégories : les méthodes d'association, les méthodes centroïdes et celles qui minimisent la variance. Dans les trois cas, la procédure de base est la même. Le processus de classification débute avec la construction d'une matrice de similarité entre les $n(n-1)/2$ couples d'individus. Nous recherchons alors dans cette matrice les deux éléments I et J les plus similaires (ou les plus proches). Ces deux derniers éléments sont alors rassemblés pour former un groupe K. La matrice de similarité est alors reconstruite et modifiée en conséquence. Nous recherchons alors de nouveau la paire d'individus (ou de groupes) la plus similaire afin de les rassembler. Le processus se déroule ainsi jusqu'à l'obtention d'une classe unique, contenant tous les éléments. Il existe plusieurs stratégies de regroupement, qui

sont en fait déterminées par la façon de définir la similarité ou dissimilarité entre les individus. Ces techniques de classification sont les techniques du plus proche voisin, du voisin le plus éloigné, de la classification par la moyenne, la méthode centroïde, la méthode de la médiane et la méthode de minimisation de la variance. Lance et William (1967) et Wishart (1969) ont montré que toutes ces méthodes satisfont la même formule de récurrence. Soient I et J deux classes de tailles respectives n_I et n_J. Ces deux classes sont fusionnées pour former la classe K, de taille $n_K = n_I + n_J$. Considérons H une troisième classe, de taille n_H. Nous notons, de façon générale, d_{IJ} la distance entre deux classes I et J et d_{ij} la distance entre deux individus i et j. Nous avons alors la relation suivante :

$$d_{HK} = \alpha_I d_{HI} + \alpha_J d_{HJ} + \beta d_{IJ} + \gamma |d_{HI} - d_{HJ}|, \tag{1.2.1}$$

où $\alpha_I, \alpha_J, \beta$ et γ sont des paramètres déterminant la nature de la méthode choisie. Il en résulte donc que la distance inter-groupe peut être calculée à partir de la distance intra-groupe (d_{IJ}) et de celle de chaque élément avec le groupe auquel il n'appartient pas.

Définition 1.2.1. *Une procédure de classification est dite procédure combinatoire si elle satisfait la relation de récurrence (1.2.1).*

Définition 1.2.2. *Une procédure de classification est dite compatible si la distance inter-groupe est la même que celle utilisée entre 2 individus.*

1.2.1. Méthodes du minimum et du maximum

1.2.1.1. *Le plus proche voisin*

La méthode du plus proche voisin, aussi appelée méthode du minimum a été introduite en 1957 par Sneath. L'équation de récurrence est la suivante :

$$d_{HK} = \frac{1}{2}d_{HI} + \frac{1}{2}d_{HJ} - \frac{1}{2}|d_{HI} - d_{HJ}|.$$

Considérons au départ N classes contenant chacune un élément. La première étape de l'algorithme revient à regrouper les deux individus séparés par la plus petite distance d_{ij}. La distance entre la nouvelle classe formée (K) et toute autre classe H est alors définie comme étant la distance entre leurs 2 éléments les plus proches. Le problème revient alors à trouver

$$d_{HK} = \min(d_{IK}, d_{JK}).$$

Par conséquent, si H et K sont regroupés ensemble, la distance entre tout individu de la classe résultante et son plus proche voisin est au plus d_{HK}. Si la distance choisie est la corrélation entre les variables, il faut donc résoudre

$$r_{HK} = \max(r_{IK}, r_{JK}),$$

où r_{HK} est le degré de similarité entre les 2 individus les plus corrélés. Il s'en suit que pour tout individu, il en existe un autre dans la même classe dont la corrélation entre eux est d'au moins r_{HK}.

Exemple 1.2.1. *Considérons 4 individus* X_1, X_2, X_3, X_4 *tels que* $X_1 = 1$, $X_2 = 1{,}5$, $X_3 = 5$, $X_4 = 6$. *Nous nous plaçons donc dans le cas* $p = 1$. *Intuitivement, nous pouvons nous attendre à obtenir un classement regroupant d'une part* X_1 *et* X_2, *et* X_3 *et* X_4 *d'autre part. Soit* D *la matrice de similarité des données calculée*

à partir de la distance euclidienne. Nous obtenons

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & & \\ 0{,}5 & 0{,}0 & & \\ 4{,}0 & 3{,}5 & 0{,}0 & \\ 5{,}0 & 4{,}5 & 1{,}0 & 0{,}0 \end{pmatrix} \end{array}. \tag{1.2.2}$$

Le plus petit élément non nul de D_1, représentant la distance entre les 2 éléments les plus proches est 0,5, ce qui correspond aux individus 1 et 2. Ces derniers sont alors rassemblés pour former une classe à 2 éléments. La distance entre cette classe et les autres éléments est calculée de la façon suivante :

$$d_{(1;2)3} = \min[d_{13}, d_{23}] = d_{23} = 3{,}5,$$
$$d_{(1;2)4} = \min[d_{14}, d_{24}] = d_{24} = 4{,}5,$$

où $d_{(i;j)k}$ représente la distance entre la classe formée par les individus i et j et l'individu k. Une nouvelle matrice peut alors être construite à partir de ces deux nouvelles distances. Nous notons $(i;j)$ la classe formée par les individus i et j. Nous obtenons :

$$D_2 = \begin{array}{c} \\ (1;2) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1;2) & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 3{,}5 & 0{,}0 & \\ 4{,}5 & 1 & 0{,}0 \end{pmatrix} \end{array}.$$

Les deux éléments les plus proches deviennent alors 3 et 4. Ces deux individus sont regroupés dans une classe à deux éléments. La distance entre les 2 classes

est donc :

$$d_{(1;2)(3;4)} = \min[d_{13}, d_{14}, d_{23}, d_{24}] = d_{23} = 3{,}5.$$

Il reste donc 2 classes (1;2) et (3;4) qui sont regroupées ensemble afin de former une classe unique contenant tous les éléments. La figure 1.2.1 illustre le résultat obtenu.

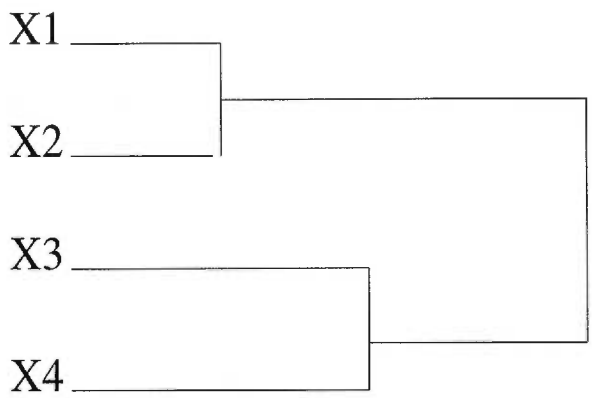

FIGURE 1.2.1. *Dendogramme obtenu après application de la méthode du plus proche voisin*

1.2.1.2. *Méthode du voisin le plus éloigné*

Comme son nom l'indique, cette méthode, aussi appelée méthode du maximum, est opposée à celle du plus proche voisin. La distance entre 2 classes est alors définie comme étant la distance entre leurs 2 éléments les plus éloignés. Nous avons alors la relation de récurrence suivante :

$$d_{HK} = \frac{1}{2}d_{HI} + \frac{1}{2}d_{HJ} + \frac{1}{2}|d_{HI} + d_{HJ}|.$$

De façon équivalente, la distance entre 2 classes est considérée comme étant le diamètre de la plus petite sphère pouvant les inclure. L'algorithme de classification

est le même que celui de la méthode précédente. La distance d_{HK} est alors :

$$d_{HK} = \max(d_{IK}, d_{JK}).$$

Dans le cas d'une mesure de similarité égale à la corrélation, nous recherchons alors

$$r_{HK} = \min(r_{IK}, d_{JK}),$$

où r_{HK} représente le degré de similarité entre les 2 individus les moins correllés.

Exemple 1.2.2. *Reprenons l'exemple 1.2.1 en appliquant cette fois la méthode du voisin le plus éloigné. À la première étape, à partir du calcul de D_1 (voir la matrice de distance donnée à l'équation (1.2.2)), les éléments 1 et 2 sont regroupés. La distance inter-groupe est cette fois ci :*

$$d_{(1;2)3} = \max[d_{13}, d_{23}] = d_{13} = 4,$$

$$d_{(1;2)4} = \max[d_{14}, d_{24}] = d_{14} = 5.$$

La matrice D_2 est reconstruite en fonction de ces nouvelles distances :

$$D_2 = \begin{array}{c} \\ (1;2) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1;2) & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 4{,}0 & 0{,}0 & \\ 5{,}0 & 1 & 0{,}0 \end{pmatrix} \end{array}.$$

Nous remarquons encore cette fois ci que la plus petite distance regroupe les individus 3 et 4. Nous obtenons alors la même classification que pour la méthode du plus proche voisin (voir figure 1.2.1).

La différence entre la méthode du plus proche voisin et celle du voisin le plus éloigné est illustré à la figure 1.2.2.

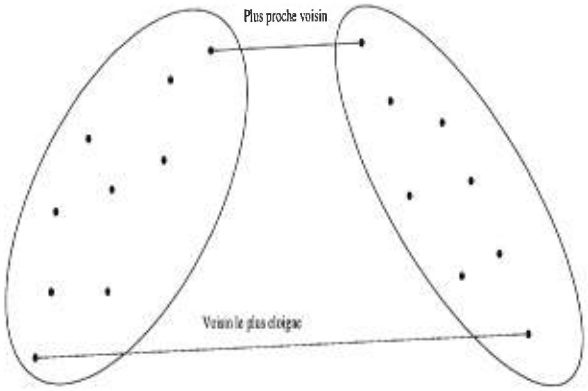

FIGURE 1.2.2. *Illustration des méthodes du plus proche voisin et du voisin le plus éloigné*

1.2.2. Méthode de la moyenne, méthode centroïde, méthode de la médiane

Ces méthodes proposées pour la première fois par Sokal et Michener en 1958 (voir Everitt, 1993) sont un compromis entre les méthodes du maximum et du minimum. La différence entre ces trois méthodes peut être vue comme une simple question de pondération.

1.2.2.1. *Méthode de la moyenne*

La distance entre 2 groupes K (formé des groupes I et J) et H est définie comme la moyenne des distances entre tous les couples X_k, X_h tels que $X_h \in H$ et $X_k \in K$. Notons d_{hi} la mesure de dissimilarité entre les éléments $h \in H$ et $i \in I$ (et non la mesure globale entre les 2 groupes). Nous avons alors :

$$\begin{aligned} d_{HK} &= \frac{1}{n_H n_K} \sum_{h \in H} \sum_{k \in K} d_{hk}, \\ &= \frac{n_I}{n_K} \frac{1}{n_H n_I} \sum_{h \in H} \sum_{i \in I} d_{hi} + \frac{n_J}{n_K} \frac{1}{n_H n_J} \sum_{h \in H} \sum_{j \in J} d_{hj}. \end{aligned}$$

Or, par définition, nous avions

$$d_{HI} = \frac{1}{n_I n_H} \sum_{i \in I} \sum_{h \in H} d_{hi}.$$

La formule de récurrence est donc la suivante :

$$\begin{aligned} d_{HK} &= \frac{n_I}{n_K} d_{HI} + \frac{n_J}{n_k} d_{HJ}, \\ &= \frac{n_I}{n_I + n_J} d_{HI} + \frac{n_J}{n_I + n_J} d_{HJ}. \end{aligned}$$

Exemple 1.2.3. *Considérons les données de l'exemple 1.2.1, ainsi que D_1 la matrice de similarité (voir l'équation (1.2.2)). La première classe formée regroupe donc, comme précédemment, les individus 1 et 2. La nouvelle matrice D_2 est donc calculée de la façon suivante :*

$$d_{(1;2)3} = \frac{1}{2}(d_{13} + d_{23}) = 3{,}75,$$
$$d_{(1;2)4} = \frac{1}{2}(d_{14} + d_{24}) = 4{,}75.$$

Nous obtenons donc :

$$D_2 = \begin{array}{c} \\ (1;2) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1;2) & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 3{,}75 & 0{,}0 & \\ 4{,}75 & 1 & 0{,}0 \end{pmatrix} \end{array}.$$

Nous pouvons constater que, comme pour les 2 méthodes précédentes, le résultat reste inchangé : les individus 3 et 4 sont regroupés ensemble à la deuxième étape.

La figure 1.2.3 est une illustration de la distance utilisée dans le contexte de la méthode de la moyenne.

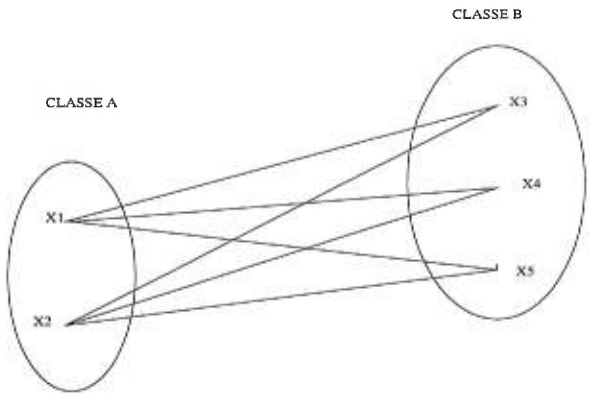

FIGURE 1.2.3. *Illustration de la distance inter-groupe : méthode de la moyenne*

1.2.2.2. *Méthode centroïde*

Dans cette approche, la distance entre deux groupes est définie comme étant la distance entre leurs deux centroïdes respectifs. Nous pouvons remarquer que la seule mesure de dissimilarité pour laquelle cette méthode est une procédure combinatoire, au sens de la relation de récurrence 1.2.1, est le carré de la distance euclidienne. Notons x_I le centroïde du groupe I. Nous avons alors :

$$x_K = \frac{n_I x_I + n_J x_J}{n_I + n_J}.$$

Plus particulièrement, si d_{HK} représente le carré de la distance euclidienne entre les groupes H et K, nous avons

$$d_{HK} = \left(x_H - \frac{n_I x_I + n_J x_J}{n_K} \right)^2.$$

La relation de récurrence est alors la suivante :

$$d_{HK} = \frac{n_I}{n_K} d_{HI} + \frac{n_J}{n_K} d_{HJ} - \frac{n_I n_J}{n_K^2} d_{IJ}.$$

Exemple 1.2.4. *Dans le cas de l'exemple 1.2.1, nous commençons par regrouper les individus 1 et 2. Le centroïde de la classe ainsi formé prend alors la valeur*

$x_{(1;2)} = (X_1 + X_2)/2$, *c'est-à-dire* $x_{(1;2)} = 1{,}25$. *La matrice* D_2 *est alors calculée comme suit :*

$$d_{(1;2)3} = d(x_{(1;2)}, X_3) = 3{,}125,$$
$$d_{(12)4} = d(x_{(1;2)}, X_4) = 3{,}625.$$

Donc nous obtenons :

$$D_2 = \begin{array}{c} \\ (1;2) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1;2) & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 3{,}125 & 0{,}0 & \\ 3{,}625 & 1 & 0{,}0 \end{pmatrix} \end{array}.$$

Les résultats ne diffèrent donc pas des autres méthodes : les individus 3 et 4 sont regroupés à la deuxième étape.

1.2.2.3. *Méthode de la médiane*

Cette méthode a l'avantage de pallier une difficulté que la méthode centroïde ne peut résoudre. Lorsque les tailles des 2 groupes (n_I et n_J) sont significativement différentes, le centroïde du groupe K, résultant de la fusion entre I et J a tendance à être très proche du centroïde du groupe le plus important. Les caractéristiques de la plus petite classe sont alors complètement absorbées par la plus grande. La méthode de la médiane ne tient pas compte de la taille des classes qu'elle rassemble. Le terme de médiane est utilisé ici, car lorsque les classes I et J sont regroupées ensemble, la distance à un troisième groupe H se trouve être le long de la médiane du triangle formé par les trois centroïdes des groupes I, J, K. Dans ce cas, nous définissons une médiane comme étant l'unique point de rencontre des trois ensembles de droites, chacune d'elle partant d'un sommet du triangle

(centroïde du groupe) et aboutissant au milieu du coté opposé. Nous obtenons :

$$d_{HK} = \frac{1}{2}d_{HI} + \frac{1}{2}d_{hj} - \frac{1}{4}d_{IJ}.$$

Exemple 1.2.5. *La matrice D_1 étant donnée (voir l'équation (1.2.2)), les nouvelles distances peuvent être aisément calculées grâce à la formule de récurrence. Nous avons :*

$$d_{(1;2)3} = 0{,}5d_{13} + 0{,}5d_{23} - 0{,}25d_{12} = 3{,}625,$$

$$d_{(1;2)4} = 0{,}5d_{14} + 0{,}5d_{24} - 0{,}25d_{12} = 4{,}625.$$

Nous obtenons :

$$D_2 = \begin{array}{c} \\ (1;2) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1;2) & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}0 & & \\ 3{,}625 & 0{,}0 & \\ 4{,}625 & 1 & 0{,}0 \end{pmatrix} \end{array}.$$

Là encore, nous ne constatons aucune différence avec les méthodes précédentes : les individus 3 et 4 sont regroupés dans la même classe.

1.2.3. Classification de Ward

Ward (1963) proposa une méthode de classification qui associe un coût à chaque partition formée. Ce coût est choisi de telle façon qu'il soit interprétable facilement. Il s'exprime comme la somme des carrés des erreurs, que nous notons ESS. En effet, nous avons :

$$ESS = \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

De la même façon, le coût associé à un groupe A est alors :

$$ESS_A = \sum_{i \in A} (X_i - \bar{X}_A)^2.$$

Nous procédons de la façon suivante. Nous débutons avec n groupes, ce qui correspond à un coût nul, pour terminer la classification avec un seul groupe contenant toutes les observations. À chaque étape, le nombre de groupes décroît de 1. Supposons k groupes formés, formant une partition optimale P_k. Nous cherchons donc la partition optimum (au niveau du coût) qui contiendrait $k-1$ groupes. Pour ce faire, nous étudions les $k(k-1)$ partitions possibles, telles que chacune de ces partitions provienne de l'union de deux éléments de la partition P_k. Le choix se porte donc sur la partition ayant le nombre de classes voulues $(k-1)$ et le coût minimal.

Cette méthode a cependant un gros désavantage par rapport aux méthodes précédentes : examiner à chaque étape toutes les partitions possibles afin de trouver la partition qui minimise le coût est très fastidieux en terme de temps de calcul.

Exemple 1.2.6. *Considérons les observations de l'exemple 1.2.1. Nous débutons donc l'algorithme en considérant 4 classes, contenant chacune une observation. La deuxième étape consiste à partitionner les 4 individus en 3 groupes. Il existe donc 6 partitions différentes :* $P_1 = \{(X_1; X_2)(X_3)(X_4)\}$, $P_2 = \{(X_1; X_3)(X_2)(X_4)\}$, $P_3 = \{(X_1; X_4)(X_2)(X_3)\}$, $P_4 = \{(X_2; X_3)(X_1)(X_4)\}$, $P_5 = \{(X_4; X_2)(X_3)(X_1)\}$ *et* $P_6 = \{(X_3; X_4)(X_2)(X_1)\}$.

Notons ESS_i *le coût du* i^e *groupe,* ESS_{Pi} *le coût calculé pour la partition* i *et* $\bar{X}_i$ *la moyenne du* i^e *groupe. Nous obtenons alors dans le premier cas :*

$$\begin{aligned} ESS_{P1} &= ESS_1 + ESS_2 + ESS_3, \\ &= \sum_{i=1,2} (X_i - \bar{X}_1)^2 + 0 + 0, \\ &= 0{,}125. \end{aligned}$$

De la même façon, nous obtenons pour les 5 autres partitions :

$$ESS_{P2} = 8,$$

$$ESS_{P3} = 12{,}5,$$

$$ESS_{P4} = 6{,}125,$$

$$ESS_{P5} = 10{,}125,$$

$$ESS_{P6} = 0{,}5.$$

La partition qui minimise le coût est la première partition, qui regroupe donc les observations 1 et 2. Cette classe est maintenant considérée comme un élément, au même titre que les individus 3 et 4. Nous cherchons donc maintenant à partitionner nos quatre observations en 2 groupes, en étudiant toutes les paires possibles. Il existe donc 3 possiblités : $P_1 = \{(X_1; X_2; X_3)(X_4)\}$, $P_2 = \{(X_1; X_2; X_4)(X_3)\}$ *et* $P_3 = \{(X_1; X_2)(X_3; X_4)\}$. *Nous calculons alors :*

$$ESS_{P1} = 9{,}5,$$

$$ESS_{P2} = 15{,}16,$$

$$ESS_{P3} = 0{,}625.$$

Nous pouvons constater que cet algorithme regroupe à nouveau les individus 3 et 4. Les résultats ne diffèrent donc pas des méthodes précédentes.

1.3. Méthodes d'optimisation

Dans cette section, nous étudions un ensemble de méthodes de classification qui produisent une partition lorsque le nombre de groupes est fixé et qui minimisent ou maximisent un critère de sélection.

1.3.1. Choix du critère de sélection

Soit g le nombre de groupes et soit n le nombre d'individus contenant chacun p variables. Posons aussi n_i le nombre d'éléments composant le i^e groupe ($i = 1, ..., g$) pour une partition donnée. L'idée de base des méthodes d'optimisation est d'associer à chaque partition des n individus, un critère $f(n, g)$ indiquant la " qualité " de celle-ci. Les partitions peuvent alors être facilement comparées entre elles. Beaucoup de critères différents ont été proposés et considérés, mais la plupart d'entre eux utilisent les trois matrices suivantes. Pour une partition donnée, nous avons alors :

$$\begin{cases} T = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})', \\ W = \frac{1}{n-g} \sum_{i=1}^{g} W_i, \\ B = \sum_{i=1}^{g} n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})', \end{cases}$$

où $W_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$ et X' représente la transposée de X, $\bar{X}$ la moyenne totale des individus et $\bar{X}_i$ la moyenne du groupe i.
Ces trois matrices, de dimension $p \times p$, représentent respectivement la dispersion totale, la dispersion intra-groupe et la dispersion inter-groupe. Elles satisfont l'équation suivante :

$$T = \frac{(n-g)}{n} W + \frac{(g-1)}{n} B.$$

Pour $p = 1$, cette équation représente une relation entre des scalaires : T représente alors la somme des carrés totaux, W la somme des carrés à l'intérieur des groupes et B la somme des carrés inter-groupes. L'équation ci-dessus peut alors être identifiée comme étant une équation d'analyse de variance à un facteur. Un critère naturel de choix de partition serait alors de choisir la partition correspondant au minimum de la somme des carrés intra-groupes, ce qui correspond aussi

de façon équivalente à la valeur maximum de la somme inter-groupes.
Lorsque $p \neq 1$, plusieurs alternatives ont été étudiées.

1.3.1.1. *Minimisation de la trace de W*

Posons $C_1 = tr(W)$. Une première alternative, proposée par Singleton et Karitz (1965) suggère de minimiser la trace de W. Nous pouvons montrer que ceci est équivalent à minimiser la somme des carrés des distances euclidiennes entre les individus et la moyenne de leur classe. Nous obtenons ainsi :

$$C_1 = \sum_{i=1}^{n} d^2_{i,c(i)}$$

où $c(i)$ représente la classe à laquelle i appartient et $d_{i,c(i)}$ est la distance euclidienne entre la i^e observation et la moyenne de la classe $c(i)$. Nous remarquons aussi que minimiser la trace de W est complètement équivalent à maximiser celle de B.

1.3.1.2. *Minimisation du déterminant de W*

Posons $C_2 = det(W)$. En analyse de variance multivariée, un des tests concernant la différence inter-groupes est basé sur le ratio des déterminants des matrices de dispersion intra-groupe et totale (voir Krzanowski, 1988). Des valeurs élevées de $det(T)/det(W)$ indiquent que la moyenne des groupes est différente. Nous pourrions alors considérer comme critère de choix de partition, la maximisation de ce quotient. Or, puisque pour toutes les partitions des n individus en g groupes, T est invariante, il suffit donc de minimiser $det(W)$.

1.3.1.3. *Maximisation de la trace de BW^{-1}*

Posons $C_3 = tr(BW^{-1})$. Un autre critère de partitionnement consiste à maximiser la trace de la matrice résultant du produit de la matrice de dispersion

inter-groupe avec l'inverse de la dispersion intra-groupe. Cette matrice, BW^{-1}, est aussi utilisée en analyse de variance multivariée. Nous pouvons remarquer que les deux derniers critères, $det(W)$ et $tr(BW^{-1})$, peuvent être exprimés en fonction des valeurs propres, λ_i, de BW^{-1} (voir Everitt (1993) p.93). Nous avons ainsi :

$$\begin{cases} Trace(BW^{-1}) = \sum_{i=1}^{p} \lambda_i, \\ \dfrac{det(T)}{det(W)} = \prod_{i=1}^{p}(1+\lambda_i). \end{cases}$$

Remarquons que les 3 derniers critères présentés ont fortement tendance à former des classes de tailles et de formes similaires. Nous pouvons alors montrer que le critère suivant, généralisation du déterminant de W, corrige cette situation en permettant la formation de groupes de tailles et de formes différentes. Nous avons :

$$C_4 = \prod_{i=1}^{g} |W_i|^{n_i}.$$

Nous verrons dans la section 2.2 comment Symons (1981) modifie le deuxième critère et l'optimise à l'aide d'arguments bayésiens.

1.3.2. Algorithme

Une fois le critère de sélection choisi, un problème numérique se pose très vite. Nous avons vu que le but des méthodes d'optimisation est de maximiser ce dernier critère. Or, le nombre de partitions possibles devient vite trop grand pour pouvoir toutes les calculer, même lorsque le nombre d'individus n est raisonnable. Soit $N(n, g)$, le nombre de partitions distinctes classant n individus en g groupes. Nous obtenons :

$$N(n,g) = \frac{1}{g}\sum_{i=0}^{g}(-1)^{g-i}\binom{n}{i}.$$

Quelques exemples, tirés de Spath(1980) illustrent cette équation :

$$\begin{cases} N(15,3) = 2375101, \\ N(20,4) = 45232115901, \\ N(100,5) = 10^{68}. \end{cases}$$

Nous remarquons donc qu'il est absolument impossible d'examiner toutes les partitions possibles des n individus en g groupes. L'algorithme de classification proposé est alors le suivant (voir Mariott, 1982) :

a) trouver une partition initiale de n individus en g groupes,

b) calculer la variation du critère, produite par chaque individu, lors du passage de leur classe initiale à une autre classe,

c) effectuer le changement qui produit la " meilleure " variation possible (grande ou petite, selon si nous cherchons à maximiser ou à minimiser le critère),

d) répéter les étapes b) et c) jusqu'à ce qu'on ne puissions plus améliorer le critère.

Le choix d'une partition de départ peut être motivé par une certaine idée *a priori* de la configuration ou peut tout simplement être le résultat d'une autre méthode de classification, hiérarchique par exemple. Bien sur, différentes partitions initiales peuvent mener à différents minimums ou maximums locaux du critère et il existe des méthodes, que nous ne détaillons pas dans ce mémoire qui permettent de trouver " la meilleure " partition initiale (voir Lorr, 1983, p. 71).

1.3.3. Variation du critère avec l'ajout d'une observation

Mariott (1982) s'est intéressé à l'effet sur le critère de choix, de l'addition d'un point X au i^{e} groupe, de moyenne $\overline{X}_i$ et de taille n_i. Cette addition modifie

W_i en W_i^* tel que

$$W_i^* = W_i + d_i d_i',$$

où $d_i = (X - \overline{X}_i)(\frac{n_i}{n_i + 1})^{1/2}$.

Nous avons :

$$tr(W_i^*) = tr(W_i) + d_i' d_i,$$

$$|W_i^*| = |W_i|(1 + d_i' W_i^{-1} d_i), \quad \text{si } |W_i| \neq 0,$$

$$W_i^* = \frac{W_i^{-1}(I - d_i d_i' W_i^{-1})}{1 + d_i' W_i^{-1} d_i}, \quad \text{si } |W_i| \neq 0.$$

Notons que ces équations restent valides lorsque nous remplaçons W_i par W. Le critère C_i de choix de partition (voir sous section 1.3.1) devient alors C_i^*. Mariott (1982) détermine la modification, pour tous critères précédents. Nous obtenons :

$$\begin{cases} C_1^* = C_1 + d_i' d_i, \\ C_2^* = C_2(1 + d_i' W^{-1} d_i), \\ C_3^* = C_3 - \dfrac{(d_i' W^{-2} d_i)}{1 + d_i' W^{-1} d_i}, \\ C_4^* = C_4 \times (1 + d_i' W^{-1} d_i)^{n_i + 1} |W_i|. \end{cases}$$

Exemple 1.3.1. *En considérant les données de l'exemple 1.2.1, nous pouvons calculer les 3 matrices T, W et B, qui dans le cas $p = 1$ se trouvent être des scalaires. Posons $g = 2$, le nombre de groupes. Choisissons la partition suivante :* $\{(X_1; X_4)(X_2; X_3)\}$. *Nous obtenons alors :*

$$T = 4{,}67,$$

$$W = \frac{1}{n - g}(W_1 + W_2) = \frac{1}{2}(12{,}5 + 6{,}125) = 9{,}31,$$

$$B = 0{,}0625.$$

Prenons par exemple le critère consistant à minimiser la trace de W. Nous avons alors $C_1 = 9{,}31$. Il s'agit maintenant de calculer, pour chaque transfert de classe de chaque individu, la variation du critère C_1. Nous devons donc étudier les 4 cas : transfert de X_1 de la classe 1 à la classe 2, transfert de X_4 de la classe 1 à la classe 2, transfert de X_2 de la classe 2 à la classe 1 et transfert de X_3 de la classe 2 à la classe 1. Dans le premier cas, nous obtenons :

$$W_2^* = 9{,}5,$$

$$W_1^* = 0.$$

Il s'en suit donc que

$$W^* = \frac{1}{n-g}(W_1^* + W_2^*) = 4,75.$$

Dans les autres cas, les résultats sont les suivants :

$$\textit{Cas 2 :}\ W^* = 5{,}58,$$

$$\textit{Cas 3 :}\ W^* = 7{,}58,$$

$$\textit{Cas 4 :}\ W^* = 7{,}00.$$

Le critère est donc minimisé (localement) dans le premier cas. Nous obtenons la partition suivante : $\{(X_4)(X_1; X_2; X_3)\}$. Si nous réitérons cette étape encore une fois, nous constatons alors que la meilleure partition possible est $\{(X_4; X_3)(X_1; X_2)\}$, ce qui correspond à un transfert de X_3 de la classe 2 à la classe 1. Nous obtenons alors $W = 0{,}31$, qui est un minimum global. Dans le cadre de notre exemple, les résultats sont donc tous identiques, quelle que soit la méthode utilisée.

1.4. Méthode du maximum de vraisemblance

Une approche très intéressante de la classification suppose les observations normales à l'intérieur de chacun des groupes. Ceci nous amène donc à considérer

un modèle mixte (voir Symons, 1981). Soit G le nombre de composantes dans le modèle mixte (qui est équivalent au nombre de groupes). La densité de chacun des X_i $(i = 1, ..., n)$ se présente alors sous cette forme :

$$f(x_i|\underset{\sim}{\theta_g}) = \sum_{g=1}^{G} \lambda_g N_p(x_i|\mu_g, \Sigma_g), \tag{1.4.1}$$

où $\underset{\sim}{\theta_g} = (\lambda_g, \mu_g, \Sigma_g)$ et où λ_g représente le paramètre de mélange et où $N_p(x_i|\mu_g, \Sigma_g)$ signifie que X_i est distribué selon une loi normale de moyenne μ_g et de matrice de covariance Σ_g. Les conditions sur le paramètre de mélange sont les suivantes :

$$\begin{cases} \sum_{g=1}^{G} \lambda_g = 1, \\ \lambda_g \geq 0, \ \forall g = 1, ..., G. \end{cases}$$

L'équation (1.4.1) est donc équivalente à supposer que X_i provient de la g^e classe avec probabilité λ_g. Nous notons par z_i la classe d'origine de X_i. Le problème revient donc à estimer les n valeurs de z_i. Posons $\mathbf{X} = (X_1, X_2, ..., X_n)$ la matrice regroupant les n individus et notons par $\underset{\sim}{\theta}$, le vecteur des paramètres $(\lambda_1, ..., \lambda_G, \mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G)$. La fonction de vraisemblance de $\mathbf{X}$ est alors calculée de la façon suivante :

$$\begin{aligned} L(\mathbf{X}|\underset{\sim}{\theta}, z) &= \left[\prod_{g=1}^{G} (\lambda_g^{n_g} |\Sigma_g|^{-n_g/2})\right] \\ &\quad \times \exp\left(-\frac{1}{2} \sum_{g=1}^{G} \sum_{C_g} (x_i - \mu_g)' \Sigma_g^{-1} (x_i - \mu_g)\right), \end{aligned} \tag{1.4.2}$$

où C_g est l'ensemble des observations appartenant au groupe g et où n_g est le nombre d'observations contenues dans C_g. Nous pouvons remarquer que pour chaque observation, G^n allocations sont possibles. Les procédures appliquant cette méthodes se doivent donc d'être optimales au niveau du temps de calcul. Le maximum de vraisemblance détermine donc $\hat{z}$, estimateur de z qui fournit l'allocation

optimale. Les paramètres dérivant de cette estimation sont alors les suivants :

$$\begin{cases} \hat{\lambda_g} = n_g/n, \\ \hat{\mu_g} = \bar{X}_g = \frac{1}{n_g} \sum_{C_g} X_i, \\ \hat{\Sigma}_g = \frac{1}{n_g} \sum_{C_g} (X_i - \bar{X}_g)(X_i - \bar{X}_g)', \end{cases}$$

pour $g = 1, ..., G$. Cette méthode nous servira de base lors du développement d'une approche bayésienne utilisant elle aussi les modèles mixtes, que nous verrons au deuxième chapitre.

Exemple 1.4.1. *Reprenons une dernière fois notre petit exemple et considérons le nombre de groupes, G, égal à 2. Dans ce cas, il existe 7 partitions possibles des 4 éléments en 2 groupes :* $P_1 = \{(X_1; X_2)(X_3; X_4)\}$, $P_2 = \{(X_1; X_3)(X_2; X_4)\}$, $P_3 = \{(X_1; X_4)(X_3; X_2)\}$, $P_4 = \{(X_1)(X_2; X_3; X_4)\}$, $P_5 = \{(X_2)(X_1; X_3; X_4)\}$, $P_6 = \{(X_3)(X_2; X_1; X_4)\}$, $P_7 = \{(X_4)(X_2; X_3; X_1)\}$. *Pour chacune de ces partitions, nous calculons les paramètres* $\hat{\lambda}_i$, $\hat{\mu}_i$ *et* $\hat{\Sigma}_i$ *(*$i = 1, 2$*). Nous calculons ensuite les vraisemblances correspondantes et nous choisissons la partition dont la vraisemblance est maximum. Par exemple, pour la partition* $\{(X_1; X_2)(X_3; X_4)\}$, *nous avons :*

$$\begin{cases} \hat{\lambda}_1 = 0{,}5, \ \hat{\lambda}_2 = 0{,}5, \\ \hat{\mu}_1 = 1{,}25, \ \hat{\mu}_2 = 5{,}5, \\ \hat{\Sigma}_1 = 0{,}125, \ \hat{\Sigma}_2 = 0{,}5, \\ L_1(\mathbf{X}|\hat{\theta}, z) = L_1 = 2{,}72. \end{cases}$$

Nous obtenons de la même façon :

$$\begin{cases} L_2 = 0{,}022, \\ L_3 = 0{,}024, \\ L_4 = 1, \\ L_5 = 1, \\ L_6 = 1, \\ L_7 = 1. \end{cases}$$

La partition rendant la vraisemblance maximale est donc la première partition, ce qui correspond donc aux résultats obtenus avec les autres méthodes.

Toutes les méthodes que nous avons présentées seront appliquées à un jeu de données dans le dernier chapitre. Elles seront comparées entre elles, puis comparées aux méthodes bayésiennes usuelles, comme les modèles mixtes, et enfin, à notre propre approche de la classification bayésienne.

Chapitre 2

MÉTHODES BAYÉSIENNES

Ce chapitre pose les bases de la théorie bayésienne ; ceci nous amène à développer deux méthodes non fréquentistes, après avoir fait une brève revue des lois multivariées usuelles en classification. La première est une extension de la méthode du maximum de vraisemblance traitée à la section 1.4. La deuxième méthode, équivalente aux méthodes d'optimisation de la section 1.3, propose une justification bayésienne intéressante quand aux critères utilisés.

2.1. CONCEPTS DE BASE DE LA THÉORIE BAYÉSIENNE

L'approche bayésienne diffère conceptuellement de l'approche classique. Considérons un modèle statistique paramétrisé par un vecteur $\underset{\sim}{\theta}$. Les méthodes statistiques classiques permettent de conduire, à partir de l'information échantillonnale, une inférence sur $\underset{\sim}{\theta}$. L'approche bayésienne consiste plutôt à combiner l'information apportée par les observations, et celle disponible *a priori* sur le paramètre, qui n'est plus seulement inconnu, mais aussi aléatoire. Nous introduisons dans cette section la définition et les propriétés du modèle statistique bayésien, les approches utilisées dans la théorie des tests et la sélection de modèles. Toutes les définitions, théorèmes et propriétés décrites dans cette section proviennent du livre de Robert (1992) et seront reprises et appliquées lors du développement de notre propre méthode de classification au chapitre 3.

2.1.1. Le modèle statistique bayésien et lois *a priori*

Considérons un modèle statistique paramétrique consistant en l'observation d'une variable aléatoire $\underset{\sim}{X}$ de densité $f(\underset{\sim}{x} \mid \underset{\sim}{\theta})$, où $\underset{\sim}{\theta}$, élément d'un sous-ensemble Θ d'un espace vectoriel de dimension finie, est inconnu. La théorie bayésienne introduit une loi sur $\underset{\sim}{\theta}$, que nous notons $\pi(\underset{\sim}{\theta})$ et que nous appelons loi *a priori* . Cette loi représente l'information disponible sur le paramètre avant l'observation de nos données.

Définition 2.1.1. *Nous appelons modèle statistique bayésien la donnée d'un modèle statistique paramétré* $f(\underset{\sim}{x} \mid \underset{\sim}{\theta})$ *et d'une loi* a priori *sur le paramètre* $\underset{\sim}{\theta}$.

En termes de probabilités le paradigme bayésien est illustré parfaitement par le théorème de Bayes.

Théorème 2.1.1. *Si* A *et* E *sont des événements indépendants tels que* $P(E) \neq 0$ *alors*

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)},$$

où A^c *représente le complément de l'événement* A.

Le modèle statistique décrit plus haut nous permet donc de déterminer la loi *a posteriori* $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$ qui combine l'information obtenue dans la loi *a priori* et dans l'échantillon. Nous aboutissons donc à la version continue du théorème de Bayes.

Théorème 2.1.2. *Soit* $\pi(\underset{\sim}{\theta})$ *une loi* a priori *sur le vecteur des paramètres* $\underset{\sim}{\theta}$ *et* $f(x|\underset{\sim}{\theta})$ *la densité de* $\underset{\sim}{X}$. *Alors, la distribution de* $\underset{\sim}{\theta}$, *conditionnellement à* $\underset{\sim}{x}$, *appelée aussi distribution* a posteriori , *a pour densité :*

$$\pi(\underset{\sim}{\theta}|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x} \mid \underset{\sim}{\theta})\pi(\underset{\sim}{\theta})}{m(\underset{\sim}{x})},$$

où $m(\underset{\sim}{x}) = \int_\Theta f(\underset{\sim}{x} \mid \underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}$.

Nous notons ici $m(\underset{\sim}{x})$ comme étant la marginale de $\underset{\sim}{X}$. La loi de $\underset{\sim}{\theta}$ conditionnellement à $\underset{\sim}{x}$ est donc proportionnelle à la loi de $\underset{\sim}{X}$ conditionnellement à $\underset{\sim}{\theta}$, multipliée par la marginale de $\underset{\sim}{\theta}$. La distribution *a posteriori* contient donc toute l'information disponible sur le paramètre, information obtenue à partir des observations et de l'information *a priori* . Elle constitue la base de toute inférence bayésienne.
De ce modèle bayésien découle une nouvelle version de la vraisemblance, que nous appelons la vraisemblance généralisée, et que nous utiliserons fréquemment par la suite.

Définition 2.1.2. *L'estimateur du maximum de vraisemblance généralisé de* $\underset{\sim}{\theta}$ *est la valeur* $\hat{\underset{\sim}{\theta}}$ *qui maximise* $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$. *La valeur* $\hat{\underset{\sim}{\theta}}$ *se trouve donc être celle qui est la plus vraisemblable pour* $\underset{\sim}{\theta}$ *étant donnée sa loi* a priori *et l'observation de l'échantillon.*

D'autres estimateurs bayésiens ont recours à la moyenne ou à la médiane de $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$. Ces trois types d'estimateurs sont calculables relativement facilement lorsque les lois *a priori* et *a posteriori* proviennent de familles conjuguées.
En effet, une des principales difficultés reliées à l'application de la théorie bayésienne est la détermination de la loi *a priori* π. Plusieurs techniques ont été développées palliant ce problème. Nous présentons ici une approche que nous utiliserons par la suite, se situant à la frontière entre statistiques bayésiennes et fréquentistes.

Définition 2.1.3. *Une famille* $\mathcal{F}$ *de lois sur* Θ *est dite conjuguée (ou fermée par échantillonnage) si pour tout* $\pi \in \mathcal{F}$, *la loi* a posteriori , $\pi(\underset{\sim}{\theta}|\underset{\sim}{x})$ *appartient également à* $\mathcal{F}$.

Lorsque $\mathcal{F}$ est paramétrée, ce qui est le cas par la suite, le passage des lois *a priori* aux lois *a posteriori* se réduit à un changement de paramètres. Nous étudierons plus particulièrement le cas de la loi normale multidimensionnelle dans le troisième chapitre. Cette méthode présente donc un attrait tout particulier, puisque les lois *a posteriori* deviennent alors toujours calculables de façon analytique.

2.1.2. Les modèles hiérarchiques

Les modèles hiérarchiques se justifient par le fait que nous disposons rarement d'informations *a priori* suffisament précises pour définir exactement la loi *a priori*. Un des buts de l'analyse bayésienne hiérarchique est donc d'incorporer cette imprécision dans le modèle en modélisant l'information en niveaux successifs et conditionnels. En d'autres termes, nous modélisons l'incertitude sur la loi *a priori* par une loi sur les paramètres de cette loi.

Définition 2.1.4. *On appelle modèle bayésien hiérarchique un modèle statistique bayésien,* $(f(\underset{\sim}{x}|\underset{\sim}{\theta}), \pi(\underset{\sim}{\theta}))$ *où la loi* a priori $\pi(\underset{\sim}{\theta})$ *est décomposée en distributions conditionnelles* $\pi_1(\underset{\sim}{\theta}|\underset{\sim}{\theta_1}), \pi_2(\underset{\sim}{\theta_1}|\underset{\sim}{\theta_2}), ..., \pi_n(\underset{\sim}{\theta_{n-1}}|\underset{\sim}{\theta_n})$, *et en une distribution marginale* $\pi_{n+1}(\underset{\sim}{\theta_n})$. *Cette dernière est entièrement spécifiée et ne dépend donc d'aucun paramètre inconnu. Les paramètres* $\underset{\sim}{\theta_i}$ *sont appelés hyperparamètres.*

Nous pouvons remarquer qu'un modèle bayésien hiérarchique est un cas particulier de modèle bayésien. En effet, nous pouvons écrire le modèle usuel

$$\underset{\sim}{X} \sim f(\underset{\sim}{x}|\underset{\sim}{\theta}),$$

$$\underset{\sim}{\theta} \sim \pi(\underset{\sim}{\theta})$$

avec

$$\pi(\underset{\sim}{\theta}) = \int_{\Theta_1 \ldots \Theta_n} \pi_1(\underset{\sim}{\theta}|\underset{\sim}{\theta_1}) \ldots \pi_n(\underset{\sim}{\theta_{n-1}}|\underset{\sim}{\theta_n}) \pi_{n+1}(\underset{\sim}{\theta_n})\, d\underset{\sim}{\theta_1}, \ldots, d\underset{\sim}{\theta_n}.$$

Cette approche nous sera particulièrement utile au troisième chapitre lors du développement du modèle de classification.

2.1.3. Les tests d'hypothèses

L'approche bayésienne propose plusieurs types de statistiques de tests. Nous introduisons ici une statistique, le facteur de Bayes, que nous utiliserons à plusieurs reprises lors de la construction de notre propre test d'hypothèses au troisième chapitre.

Soit un modèle statistique paramétré, $f(\underset{\sim}{x}|\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \Theta$. Étant donné deux sous-ensemble $\Theta_0 \subset \Theta$ et $\Theta_1 \subset \Theta$, tels que $\Theta_0 \cap \Theta_1 = \emptyset$, nous cherchons à déterminer si $\underset{\sim}{\theta}$, le " vrai " paramètre, appartient à Θ_0, c'est-à-dire à tester l'hypothèse $H_0 : \underset{\sim}{\theta} \in \Theta_0$ versus $H_1 : \underset{\sim}{\theta} \in \Theta_1$.

Définition 2.1.5. *Nous appelons facteur de Bayes, ou encore cote de Bayes, le rapport*

$$B(\underset{\sim}{x}) = \frac{P(\underset{\sim}{\theta} \in \Theta_0|\underset{\sim}{x})}{P(\underset{\sim}{\theta} \in \Theta_1|\underset{\sim}{x})} \frac{P(\underset{\sim}{\theta} \in \Theta_1)}{P(\underset{\sim}{\theta} \in \Theta_0)}.$$

Ce rapport de probabilités évalue donc la modification de la vraisemblance relative de l'hypothèse nulle qui est due aux observations. Il se compare naturellement à 1. Si ce rapport est supérieur à 1, nous dirons que nous acceptons l'hypothèse H_0 et s'il est inférieur à 1, nous acceptons alors l'hypothèse H_1. Si ce rapport vaut 1, cela signifie donc que les deux hypothèses sont également vraisemblables. Nous pouvons remarquer que dans le cas particulier où Θ_0 est réduit à un point , $\{\theta_0\}$, et $\Theta_1 = \{\theta_1\}$, le facteur de Bayes est équivalent au rapport de

vraisemblance classique :

$$B^{\pi}(\underline{x}) = \frac{f(\underline{x}|\theta_0)}{f(\underline{x}|\theta_1)}.$$

2.1.4. Critères de sélection de modèles

La sélection de modèles est un sujet qui nous concerne tout particulièrement puisque beaucoup de méthodes bayésiennes de classification y réfèrent. Un des critères les plus utilisés en statistique classique a été introduit dans Akaike (1978) et fait référence au maximum de vraisemblance. Posons P le nombre de paramètres du modèle et $\hat{\theta}$ l'estimateur du maximum de vraisemblance. Nous pouvons alors définir le critère suivant :

$$AIC = (-2)l(\hat{\theta}) + 2P,$$

où la fonction l représente le logarithme de la vraisemblance. La procédure qui sélectionne le modèle minimisant le AIC parmi un ensemble de modèles est appelée minimisation du AIC. Nous pouvons remarquer que cette procédure pénalise assez fortement les modèles ayant un grand nombre de paramètres (en effet, plus le nombre de paramètres à estimer d'un modèle est élevé, plus l'erreur globale d'estimation est grande).
L'approche bayésienne de ce critère est la suivante (voir Akaike, 1979) :

$$BIC = (-2)l(\hat{\theta}) + (\log(N))P + C,$$

où N est le nombre d'observations et où C est une constante. L'introduction du BIC est basée sur des arguments bayésiens et fournit notamment un estimateur

cohérent du " vrai modèle ". Le concept d'estimateur cohérent en statistique bayésienne est défini ci-dessous (voir Bernardo et Smith, 1994, p.312).

Définition 2.1.6. *Soit e_n une expérience consistant en l'observation d'un échantillon aléatoire $X = \{X_1, ..., X_n\}$ de densité $f(X|\theta)$, où $X \in \mathcal{X}$ et $\theta \in \Theta \subseteq \mathbb{R}$. Posons aussi z_{kn} le résultat d'une k-répétition de e_n. S'il existe un estimateur $\hat{\theta}_{kn} = \hat{\theta}_{kn}(z_{kn})$ tel que, avec probabilité 1 :*

$$\lim_{k\to\infty} \hat{\theta}_{kn} = \theta,$$

alors cet estimateur est cohérent.

D'autre part, Akaike (1979) propose une justification bayésienne au critère du *AIC* en démontrant que cet estimateur est minimax et engendre une approximation raisonnable de l'estimateur de Bayes sous l'hypothèse d'égalité des probabilités *a priori* de chaque modèle (les définitions d'estimateur de Bayes et minimax peuvent être trouvées dans Robert (1992) à la section 2.4).

2.2. Lois multidimensionnelles usuelles

Avant de développer les méthodes bayésiennes, nous nous devons de faire une brève revue des lois multivariées fréquemment utilisées. Les définitions et propriétés citées dans cette section sont tirées du livre d'Anderson (1984).
Nous commençons par une révision de la densité de la loi normale multivariée.

Définition 2.2.1. *Si $X = \{X_1, ..., X_p\}$, un vecteur aléatoire de $\mathbb{R}^p$, a pour densité*

$$f(x|\mu, \Sigma) = (\det(\Sigma))^{-1/2}(2\pi)^{-p/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right),$$

où $\mu \in \mathbb{R}^p$ et Σ est une matrice $p \times p$ symétrique définie positive, alors x suit une loi multinormale $\mathcal{N}_p(\mu, \Sigma)$ d'espérance μ et de matrice de variance-covariance Σ.

Nous pouvons aussi écrire la densité de X sous la forme suivante équivalente :

$$f(x|\mu,\Sigma) = (\det\Sigma)^{-1/2}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}Tr[\Sigma^{-1}(x-\mu)(x-\mu)']\right),$$

où $Tr(A)$ est défini de façon générale comme étant la trace de la matrice A.

Une loi courante dans le choix de la loi *a priori* sur les paramètres de variance est la loi de Wishart, extension multivariée de la loi gamma unidimensionnelle.

Théorème 2.2.1. *Soient $Z_1, ..., Z_n$ n vecteurs indépendants et identiquement distribués selon des lois normales multivariées de moyenne 0 et de matrice de variance Σ. La densité de $A = \sum_{i=1}^{n} Z_i Z_i'$ est*

$$w(A|\Sigma,n) = \begin{cases} \dfrac{|A|^{(n-p-1)/2}\exp(-\frac{1}{2}Tr\Sigma^{-1}A)}{2^{pn/2}|\Sigma|^{n/2}\Gamma_p(n/2)}, & \text{si } A \text{ est définie positive,} \\ 0, & \text{sinon,} \end{cases}$$

où $\Gamma_p(t)$ est la fonction gamma multivariée définie comme suit :

$$\Gamma_p(t) = \pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[t-\frac{1}{2}(i-1)\right].$$

La matrice A suit donc une loi de Wishart de paramètres n et Σ, notée $A \sim W(n,\Sigma)$.

La loi de Wishart possède plusieurs propriétés intéressantes que nous énonçons dans le théorème et la proposition ci-dessous.

Théorème 2.2.2. *Si $A_1, ..., A_q$ sont des variables indépendantes et distribuées telles que $A_i \sim W(n_i,\Sigma)$ alors $A = \sum_{i=1}^{q} A_i$ est distribuée selon une loi de Wishart $W(\sum_{i=1}^{q} n_i, \Sigma)$.*

Proposition 2.2.1. *Soit A une matrice distribuée selon une loi de Wishart $W(n, \Sigma)$ et soit C une matrice $p \times p$ non singulière telles que*

$$A = CBC'.$$

Alors, la matrice B est distribuée selon une loi de Wishart $W(n, \phi)$ telle que

$$\phi = C^{-1}\Sigma(C')^{-1}.$$

Les démonstrations du théorème et de la proposition ci-dessus peuvent être trouvées dans Anderson(1984, p.254).

De la même façon que la loi inverse gamma provient de la loi gamma, la loi inverse Wishart est " extraite " de la loi Wishart.

Théorème 2.2.3. *Si A a une distribution $W(n, \phi)$, alors $B = A^{-1}$ suit une loi inverse Wishart de paramètres (n, Ψ). La densité de B est la suivante :*

$$w(B|n, \Psi) = \begin{cases} \dfrac{|\psi|^{n/2}|B|^{-(n+p+1)/2}\exp(-\dfrac{1}{2}Tr\Psi B^{-1})}{2^{np/2}\Gamma_p(n/2)}, & \text{si } B \text{ est définie positive,} \\ 0, & \text{sinon.} \end{cases}$$

où n est le degré de liberté de la loi et où $\Psi = \phi^{-1}$. Nous notons alors $B \sim IW(n, \psi)$.

Une dernière propriété très intéressante de la loi inverse Wishart est qu'elle est une famille conjuguée pour la matrice de variance-covariance d'une loi normale multivariée, comme l'illustre le théorème suivant.

Théorème 2.2.4. *Soit $A = \sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})'$ la matrice proportionnelle à celle de variance-covariance définie positive d'un échantillon $X_1, ..., X_N$ de N vecteurs indépendants de $\mathbb{R}^p$ $(N > p)$ et identiquement distribués $\mathcal{N}(\mu, \Sigma)$ (Σ définie positive). Si A est distribuée selon une loi de Wishart $W(n, \Sigma)$ et Σ suit une*

loi inverse Wishart $IW(m, \psi)$*, alors la distribution conditionnelle de* Σ *sachant* A *est une loi inverse Wishart* $IW(n+m, A+\psi)$*. Sous un tel modèle, la famille des lois inverse Wishart est donc conjuguée pour la matrice de variance-covariance* Σ*.*

DÉMONSTRATION. Soit $\pi(A, \Sigma)$ la densité conjointe de A et Σ. Nous avons donc :

$$\pi(A, \Sigma) = \pi(A|\Sigma)\pi(\Sigma).$$

Ceci implique donc que

$$\pi(A, \Sigma) = \frac{|A|^{(n-p-1)/2}|\psi|^{m/2}|\Sigma|^{-(m+n+p+1)/2}\exp\left(\frac{-1}{2}Tr(\Sigma^{-1}(A+\psi))\right)}{2^{p(n+m)/2}\Gamma_p(n/2)\Gamma_p(m/2)}.$$

De plus, la distribution marginale de A est

$$\begin{aligned} m(A) &= \int \pi(A|\Sigma)\pi(\Sigma)\, d\Sigma, \\ &= \frac{|\psi|^{m/2}|A|^{(n-p-1)/2}}{2^{(n+m)p/2}\Gamma_p(n/2)\Gamma_p(m/2)} \\ &\quad \int |\Sigma|^{-(m+n+p+1)/2}\exp\left(\frac{-1}{2}Tr(\Sigma^{-1}(A+\psi))\right) d\Sigma. \end{aligned}$$

On reconnaît facilement dans cette intégrale une loi $IW(m+n, A+\psi)$. Nous obtenons donc

$$m(A) = \frac{|A|^{(n-p-1)/2}|\psi|^{m/2}}{2^{p(n+m)/2}\Gamma_p(n/2)\Gamma_p(m/2)} \frac{2^{p(m+n)/2}\Gamma_p((m+n)/2)}{|\psi+A|^{(m+n)/2}}.$$

La distribution *a posteriori* de Σ est alors

$$\begin{aligned} \pi(\Sigma|A) &= \frac{\pi(A|\Sigma)\pi(\Sigma)}{m(A)}, \\ &= \frac{|\psi+A|^{(m+n)/2}|\Sigma|^{-(m+n+p+1)/2}\exp\left(\frac{-1}{2}Tr(\Sigma^{-1}(A+\psi))\right)}{2^{p(n+m)/2}\Gamma_p((m+n)/2)}, \end{aligned}$$

ce qui est la distribution d'une loi inverse Wishart $IW(n+m, A+\psi)$. □

2.3. Une approche bayésienne de la méthode du maximum de vraisemblance

Symons (1981) développa un critère bayésien qui provient d'une extension de la méthode du maximum de vraisemblance classique (voir section 1.4). Considérons le modèle mixte développé dans cette dernière section. L'approche bayésienne requiert donc une distribution *a priori* sur le paramètre $\underset{\sim}{\theta}$ que nous notons $\pi(\underset{\sim}{\theta})$. Souvent, la loi non informative de Jeffrey est utilisée à cette intention. Elle est basée sur la matrice d'information de Fisher définie ci-dessous.

Définition 2.3.1. *La matrice d'information de Fisher est une matrice $P \times P$ (si P est la dimension du vecteur $\underset{\sim}{\theta} \in \Theta$, où Θ est le support de $\underset{\sim}{\theta}$) dont les éléments sont notés*

$$\mathcal{I}_{ij}(\underset{\sim}{\theta}) = -\mathbb{E}_{\underset{\sim}{x}}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left(\log\left[f(\underset{\sim}{X} \mid \underset{\sim}{\theta})\right]\right)\right],$$

où $\mathbb{E}_{\underset{\sim}{x}}$ représente l'espérance calculée par rapport à la loi $f(\underset{\sim}{x}|\underset{\sim}{\theta})$ des observations. La loi non informative de Jeffrey est alors

$$\pi^*(\underset{\sim}{\theta}) = [\det(\mathcal{I}(\underset{\sim}{\theta}))]^{1/2}.$$

Exemple 2.3.1. *Calculons la loi de Jeffrey lorsque la variable X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, où $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$. Afin de calculer la matrice d'information de Fisher, nous devons d'abord calculer les quatres dérivées partielles. Nous avons :*

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

Et donc, selon la définition de l'information,

$$I(\mu,\sigma^2) = -\begin{pmatrix} \mathbb{E}_X\left[\frac{\partial^2 f(x|\mu,\sigma^2)}{\partial\mu^2}\right] & \mathbb{E}_X\left[\frac{\partial^2 f(x|\mu,\sigma^2)}{\partial\mu\partial\sigma^2}\right] \\ \mathbb{E}_X\left[\frac{\partial^2 f(x|\mu,\sigma^2)}{\partial\mu\partial\sigma^2}\right] & \mathbb{E}_X\left[\frac{\partial^2 f(x|\mu,\sigma^2)}{\partial(\sigma^2)^2}\right] \end{pmatrix}$$

Ceci revient donc à écrire

$$I(\mu,\sigma^2) = -\begin{pmatrix} \mathbb{E}_X\left[\frac{-2}{\sigma^2}\right] & \mathbb{E}_X\left[\frac{-(X-\mu)}{\sigma^4}\right] \\ \mathbb{E}_X\left[\frac{-(X-\mu)}{\sigma^4}\right] & \mathbb{E}_X\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right] \end{pmatrix}$$

Nous obtenons donc :

$$I(\mu,\sigma^2) = \begin{pmatrix} \frac{2}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^4} \end{pmatrix}$$

La loi de Jeffrey non informative a priori *pour le paramètre* (μ,σ^2) *est donc :*

$$\begin{aligned} \pi(\mu,\sigma^2) &= \sqrt{|I(\mu,\sigma^2)|}, \\ &= (\sigma^2)^{-3/2}. \end{aligned}$$

Cette loi, ne s'intégrant pas à 1 est cependant une loi impropre. En réalité, il arrive très fréquemment que ce type de loi soit impropre. En effet, favoriser les valeurs de μ *et* σ *telles que l'information soit grande revient donc bien à minimiser l'importance de l'information* a priori *et se révèle donc aussi non informatif que possible. Dans le cas de la loi normale multivariée,* $\mathcal{N}_p(\mu,\Sigma)$*, où* Σ *est une matrice définie positive, la loi* a priori *devient alors (voir Box et Tiao, 1992, section 8.2.2)*

$$\pi(\mu,\Sigma) = |\Sigma|^{-(p+1)/2}.$$

Les lois *a priori* non informatives sont utilisées lorsqu'il est impossible de bâtir une distribution *a priori* basée sur des considérations subjectives. Dans notre modèle mixte, la fonction de vraisemblance est alors

$$L(\underset{\sim}{x}|z) = \int L(\underset{\sim}{x}|\underset{\sim}{\theta}, z)\pi(\underset{\sim}{\theta})\, d\underset{\sim}{\theta}, \tag{2.3.1}$$

où $L(\underset{\sim}{x}|\underset{\sim}{\theta}, z)$ est la fonction de vraisemblance classique définie à l'équation (1.4.2) au premier chapitre. La fonction (2.3.1) représente donc une fonction de vraisemblance généralisée (voir section 2.1.1). De ce maximum de vraisemblance généralisée va découler deux critères de sélection de modèle en classification bayésienne, correspondant à deux cas spécifiques : le cas des variances égales, ce qui correspond à $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$, et le cas des variances différentes. La loi *a priori* utilisée dans le premier cas est non informative. Posons $\Sigma_1 = \Sigma_2 = ... = \Sigma_g = \Sigma$. Nous avons alors :

$$\begin{aligned}\pi(\underset{\sim}{\theta}) &= \pi_1(\lambda_1...\lambda_G)\pi_2(\mu_1...\mu_G|\Sigma)\pi_3(\Sigma) \\ &\propto (\prod_{g=1}^{G}\lambda_g)^{-1}|\Sigma|^{-(p+1)/2}.\end{aligned}$$

La loi *a priori* de Σ utilisée à l'équation précédente est telle que (voir Geisser et Cornfield, 1963, p.369) :

$$\pi_3(\Sigma) \propto |\Sigma|^{-a/2}, \tag{2.3.2}$$

où a est un paramètre d'ajustement, tel que $p + 1 \leq a \leq n$. Nous pouvons remarquer que cette dernière loi est en fait un cas limite d'une loi inverse Wishart $IW(k, \psi_m)$, où $\psi_n = \psi_0/n$, $n \to \infty$ et où $k = a-p-1$. La vraisemblance classique $L(\underset{\sim}{x}|\underset{\sim}{\theta}, z)$ (voir équation 1.4.2) et la loi *a priori* sur $\underset{\sim}{\theta}$ sont alors multipliées (voir équation 2.3.1) et intégrées par rapport à $\underset{\sim}{\theta}$. Symons (1981) montre alors que

TABLEAU 2.3.1. *Critère bayésien pour les sept partitions*

Partition	*W*	*Critère C*
$(X_1; X_2)(X_3; X_4)$	*0,3125*	*-0,97*
$(X_1; X_3)(X_2; X_4)$	*9,06*	*5,79*
$(X_1; X_4)(X_2; X_3)$	*9,31*	*5,84*
$(X_1; X_2; X_3)(X_4)$	*4,75*	*2,82*
$(X_1; X_2; X_4)(X_3)$	*7,58*	*3,76*
$(X_1; X_3; X_4)(X_2)$	*7*	*6,60*
$(X_2; X_3; X_4)(X_1)$	*6,01*	*3,29*

l'allocation bayésienne optimum, $\hat{z}$, revient à partitioner les données en G groupes tels que le critère

$$C = (n - G)\log|W| + \sum_{g=1}^{G}\{p\log(n_g) - 2\log(\Gamma(n_g))\} \tag{2.3.3}$$

est minimisé. Ici, W représente la matrice intra-groupe définie à la section 1.3. Nous pouvons remarquer que ce critère est une modification du deuxième critère développé par Singleton et Karitz (1965), sous des arguments bayésiens (voir section 1.3).

Exemple 2.3.2. *Reprenons l'exemple 1.3.1 que nous avons traité au premier chapitre dans le cadre des méthodes d'optimisation. Nous étudions toutes les partitions possibles de quatre individus en G groupes. Lorsque* $G = 2$, *par exemple, il existe sept partitions possibles. Dans chacun des sept cas, nous allons calculer le critère (2.3.3) et choisir la partition le minimisant. Les résultats sont présentés dans le tableau 2.3.1. Nous constatons donc que, lorsque* $G = 2$, *la première partition est optimale. Nous notons alors* $C^*_{G=2} = -0{,}97$. *Si nous réitérons cette*

étape pour chaque nombre de groupes possibles, nous obtenons :

$$C^*_{G=4} = 0, \text{ correspondant à } \{(X_1)(X_2)(X_3)(X_4)\},$$
$$C^*_{G=3} = -3{,}46, \text{ correspondant à } \{(X_1; X_2)(X_3)(X_4)\},$$
$$C^*_{G=1} = 1{,}51, \text{ correspondant à } \{(X_1; X_2; X_3; X_4)\}.$$

La partition optimale consiste donc à partitioner les 4 observations en 3 groupes, ce qui diffère des autres méthodes étudiées au chapitre précédent. Notons que nous n'appliquerons pas cette méthode au jeu de données décrit au dernier chapitre (iris de Fisher), le taux d'erreur étant bien trop important (supérieur à 60 %). En effet, il semble que cette méthode ne soit pas efficace dans le cas de ces données.

Dans le cas des matrices de variance-covariance non égales, la loi *a priori* sur $(\Sigma_1, ..., \Sigma_g)$ sous l'hypothèse d'indépendance des Σ_i $(i = 1, ..., g)$ est égale à

$$\pi(\Sigma_1, ..., \Sigma_g) = \prod_{g=1}^{G} \pi_4(\Sigma_g),$$

où $\pi_4(\Sigma_g)$ est identique à $\pi_3(\Sigma)$ (voir équation (2.3.2)) pour $g = 1, ..., G$. L'allocation optimale $\hat{z}$ est telle que l'équation (2.3.1) est maximisée et elle est équivalente à la partition minimisant le critère

$$\sum_{g=1}^{G} \left((n_g - 1) \log |W_g| + p \log(n_g) - p(n_g + 1) \log 2 \right.$$
$$\left. -2 \left[\log \Gamma(n_g) + \sum_{i=1}^{p} \log \Gamma \left\{ \frac{n_g + p + 1 - i}{2} \right\} \right] \right).$$

Remarquons que ce critère est considérablement plus fastidieux à évaluer que le premier.

2.4. JUSTIFICATION BAYÉSIENNE DES MÉTHODES D'OPTIMISATION CLASSIQUES

Binder (1978) développe un critère et une méthode qui, sous des arguments bayésiens, sont équivalents aux méthodes d'optimisation classiques, développées à la section 1.3. Là encore, le modèle proposé est basé sur des distributions mixtes. Posons $X_1, ..., X_n$, n vecteurs de dimension p. Supposons, en conservant les notations utilisées précédemment, que le vecteur $\underset{\sim}{z} = (z_1, ..., z_n)$ représente la partition " réelle " des données en G groupes, où $z_i = k$ signifie que X_i appartient au k^{e} groupe et où G est inconnu. Le problème revient donc à spécifier $\hat{z}$, un estimateur de z. Les valeurs de G, z et un vecteur de paramètres $\underset{\sim}{\theta}$ étant donnés, les X_i sont indépendants, de densité $h_{z_i}(\underset{\sim}{x_i}|\underset{\sim}{\theta})$ avec probabilité λ_{z_i}. Les fonctions $h_1(\underset{\sim}{x}|\underset{\sim}{\theta}), ..., h_G(\underset{\sim}{x}|\underset{\sim}{\theta})$ sont connues en termes de $\underset{\sim}{x}$ et de $\underset{\sim}{\theta}$. En d'autres termes, nous avons

$$Pr(z_i = k|\lambda, G) = \lambda_k, \ i = 1, ..., n.$$

La densité *a priori* pour les paramètres inconnus G, z, $\underset{\sim}{\theta}$ et λ est donnée par :

$$\pi(G, \lambda, z, \underset{\sim}{\theta}) = \pi_1(G)\pi_2(\lambda|G) \prod_{i=1}^{G} \lambda_i^{n_i} \pi_3(\underset{\sim}{\theta}|G, \lambda, z),$$

où n_i est le nombre d'éléments contenus dans le groupe i (notons que $\sum_{i=1}^{n} n_i = n$) et où $\pi_1(G)$ est la loi de probablité sur le nombre de groupes G. La marginale de $\underset{\sim}{X_k}$ s'écrit quand à elle de cette façon :

$$f(\underset{\sim}{x_k}|G, \lambda, \underset{\sim}{\theta}) = \sum_{i=1}^{G} \lambda_i h_i(\underset{\sim}{x_k}|\underset{\sim}{\theta}),$$

à la condition que $\pi_3(\underset{\sim}{\theta}|G, \lambda, \underset{\sim}{z})$ ne dépende pas de $\underset{\sim}{z}$. Les X_k sont alors indépendants et identiquement distribués. La probabilité *a posteriori* de $\underset{\sim}{z}$ est donnée

par

$$f(z|\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n) \propto \sum_g \pi_1(g)\pi_4(\underset{\sim}{z}|G=g) \int \pi_5(\underset{\sim}{\theta}|\underset{\sim}{z}, G=g) \prod_{i=1}^{g} \prod_{k \in A_i(z)} h_i(\underset{\sim}{x}_k|\underset{\sim}{\theta})\, d\underset{\sim}{\theta},$$

où $A_i(z)$ représente l'ensemble des indices k tels que $z_k = i$ et où $\pi_1(g)$ est la probabilité d'avoir une partition contenant g groupes.

Binder (1978) introduit alors un coût relié à l'estimation du vecteur $\underset{\sim}{z}$, noté $C(\hat{\underset{\sim}{z}}|z)$. Posons alors $\hat{\underset{\sim}{z}}$ la partition minimisant l'espérance du coût *a posteriori* , que nous notons $\mathbb{E}[C(\hat{\underset{\sim}{z}}|\underset{\sim}{z})|\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n]$. Binder énonce trois principes à respecter :

i) le paramètre $\underset{\sim}{z}$ ne doit pas dépendre de l'ordre des observations $\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n$,

ii) les indices associés à la " vraie partition " ne doivent pas interférer dans l'estimation du vecteur $\underset{\sim}{z}$,

iii) les indices de la partition estimée se doivent d'être arbitraires.

Il existe différentes fonctions de coût. Tout d'abord, étudions le coût le plus simple. Nous supposons que $C(\hat{\underset{\sim}{z}}|\underset{\sim}{z}) = 0$ dès qu'il existe une permutation v telle que $v(\hat{z}_1, ..., \hat{z}_n\} = \{z_1, ..., z_n\}$ et 1 sinon. Sous un tel coût, un estimateur $\hat{\underset{\sim}{z}}$ de $\underset{\sim}{z}$ correspond au $\underset{\sim}{z}$ qui maximise la probabilité *a posteriori* .

Binder (1978) développe ensuite le coût linéaire tel que $C(\hat{\underset{\sim}{z}}|\underset{\sim}{z}) = \sum_{i,j} c_{ij} n_{ij}$ où n_{ij} est défini comme étant le nombre d'observations appartenant en réalité au i^{e} groupe et classées dans le j^{e} groupe. De façon générale, nous avons $c_{ij} > c_{jj}$. Ce coût est équivalent à perdre c_{ij} lorsqu'une des observations provenant de la j^{e} classe est assignée au i^{e} groupe. Sous ce coût, l'estimateur $\hat{\underset{\sim}{z}}$ est tel que $\hat{\underset{\sim}{z}} = i^*$ si

$$\sum_j (c_{ij} - c_{jj}) Pr(z_k = j|\underset{\sim}{x}_1, ..., \underset{\sim}{x}_n)$$

est minimisé en $i = i^*$. En réalité, telle que présentée ci-dessus, cette fonction de coût satisfait uniquement le principe i). Or, si nous lui " imposons " le respect du

deuxième principe, nous obtenons $c_{ij} = c_{ii}\ \forall i$ et j. Sous cette restriction, nous aboutissons à $\hat{z}_k = i^*$ pour chaque indice k pour lequel c_{ii} est minimum en i^*. Cet estimateur ne dépend donc plus des données, ce qui n'a plus de sens. La vérification du troisième principe implique alors que $c_{ij} = c_{jj}$ et donc, chaque partition engendre le même coût, que l'on ait fait une erreur d'estimation ou non. Pour toutes ces raisons, les applications du coût linéaire sont très restreintes.

Enfin, afin d'introduire le coût quadratique, nous devons au préalable introduire certaines notations. Nous développons ce coût uniquement dans le cas où chacun des trois principes énoncés plus haut sont respectés. Supposons une perte de b_{rt} lorsque $\hat{z}_k = r$, $\hat{z}_l = t$ et que en réalité $z_k = z_l$. De la même façon, nous supposons une perte de c_{rt} si $\hat{z}_k = r$, $\hat{z}_l = t$ alors qu'en réalité $z_k \neq z_l$. Enfin, notons n_{rs} le nombre d'observations classées dans le groupe r et appartenant en réalité au groupe s. Posons aussi :

$$b_{rt} = \begin{cases} b_1 \text{ si } r = t, \\ b_2 \text{ sinon,} \end{cases}$$

pour tous r et t. De même, nous définissons

$$c_{rt} = \begin{cases} c_1 \text{ si } r = t, \\ c_2 \text{ sinon.} \end{cases}$$

Le coût C est alors défini de la façon suivante :

$$C(\hat{\underline{z}}|\underline{z}) = \frac{1}{2}\Big\{(b_2 - b_1 + c_2 - c_1)\sum n_{rs}^2 + (c_1 - c_2)\sum \hat{n}_r^2 + (b_2 - c_2)\sum n_s^2 + c_2 n^2 - b_1 n\Big\},$$

où n_s est le nombre d'individus classés dans le s^{e} groupe ($n_s = \sum_r n_{rs}$) et où $\hat{n}_r$ est un estimé de n_r tel que $\hat{n}_r = \sum_s n_{rs}$. Cette fonction de coût engendre donc une perte de $b_2 - b_1$ si deux observations appartenant initialement au même groupe

sont classées dans deux groupes différents, et une perte de $c_1 - c_2$ si deux observations appartenant initialement à deux groupes différents sont classées ensemble. Nous pouvons voir le rapport $(b_2 - b_1)/(c_1 - c_2)$ comme une mesure relative de l'importance de la cohésion interne par rapport à la cohésion externe. Ces deux derniers concepts sont définis dans Cormack (1971, p.329). En effet, certaines méthodes de classification privilégient l'isolation externe, en se basant par exemple sur la restriction maximale acceptable pour laquelle deux individus sont séparés dans deux classes différentes. D'autres méthodes, au contraire, privilégient la cohésion interne en se basant sur la plus petite corrélation possible entre deux individus, afin que ceux-ci soient classés ensemble.

Lorsque les fonction h_i représentent la densité d'une loi normale multivariée de moyenne μ_i et de matrice de variance Σ_i supposées égales entre elles $(\Sigma_1 = ... = \Sigma_G = \Sigma)$, la loi de μ_i $(i = 1, ..., G)$ est donnée comme étant une loi normale multivariée (Σ, $\underset{\sim}{z}$ et G étant connus), de moyenne ν_i et de matrice de variance $\alpha_i\Sigma$. Ce choix est tout à fait justifiable par le fait que la loi normale est une loi conjuguée pour l'espérance d'une loi normale. Nous considérons trois cas différents : $\Sigma = \sigma^2 I$, Σ =diag$(\sigma_1^2, ..., \sigma_p^2)$ et enfin, le cas ou Σ est entièrement non spécifiée. Les lois *a priori* pour $\Sigma|G, \underset{\sim}{z}$ sont alors les suivantes :

Cas 1 : $\sigma^{-2} \sim$ Gamma (ρ, τ^{-1}),

cas 2 : $\sigma_i^{-2} \sim$ Gamma (ρ_i, τ_i^{-1}) $i = 1, ..., G$ et les σ_i sont indépendants,

Cas 3 : $\Sigma^{-1} \sim$ Wishart (ρ, V^{-1}),

où les valeurs de ρ, ρ_i, τ, τ_i et V sont connues. Supposons $\alpha_1 = ... = \alpha_G = \alpha$, c'est-à-dire que nous supposons les paramètres (α, ρ, τ) indépendants de $\underset{\sim}{z}$. La loi

a posteriori de $\underset{\sim}{z}$ sachant G est alors proportionnelle à

$$\pi_4(\underset{\sim}{z}|G)\prod_{i=1}^{G} n_i^{p/2}\left\{\tau+\frac{1}{2}Tr(W)\right\}^{(\rho+np)/2)},$$

où n_i est le nombre d'éléments composant le groupe i et W est la matrice intra-groupe définie à la section 1.3. De plus, si

$$\pi(\underset{\sim}{z}|G) \propto \prod_{i=1}^{G} n_i^{p/2},$$

alors, la probabilité *a posteriori* de z est maximisée lorsque la trace de W est minimisée. Dans le second cas, si nous supposons $(\alpha_i, \rho_i, \tau_i) = (\alpha, \rho, \tau)$, c'est-à-dire que nous supposons qu'aucun des paramètres ne dépend de G et de $\underset{\sim}{z}$ et si de plus, nous supposons α assez grand et τ assez petit, la probabilité *a posteriori* de $\underset{\sim}{z}$, G étant donné, est maximisée lorsque $|diag(w_{\cdot}^{(1,1)}, ..., w_{\cdot}^{(p,p)})|$ est minimisé, en considérant les notations suivantes :

$$W = \sum_{i=1}^{G} W_i,$$

$w_i^{(j,k)}$ est l'élément (j,k) de la matrice W_i,

$$w_{\cdot}^{(i,j)} = \sum_{k} w_i^{(j,k)}.$$

Enfin, dans le dernier cas, si α est assez grand, si (α, ρ, ν) ne dépend pas de $\underset{\sim}{z}$ et si V est proche de la matrice identiquement nulle, la probabilité *a posteriori* de $\underset{\sim}{z}$ sachant G est maximisée lorsque $det(W)$ est minimisé. Dans les deux derniers cas, les probabilités *a posteriori* sont données dans Binder (1978). Ces résultats nous fournissent donc une justification bayésienne aux trois méthodes d'optimisation présentées à la section 1.3. Bien sûr, nous ne traitons pas d'exemple de cette méthode, celle-ci étant identique aux méthodes d'optimisation.

Nous avons donc introduit certaines méthodes bayésiennes de classification, que nous comparerons plus tard, au dernier chapitre, à notre propre méthode que nous développons au prochain chapitre. Cette méthode utilise de nombreuses notions que nous avons déjà présentées.

Chapitre 3

CLASSIFICATION ET TESTS D'HYPOTHÈSES

Dans les deux derniers chapitres, nous avons fait une revue des méthodes usuelles de classification, d'abord d'un point de vue classique, puis d'un point de vue bayésien. La méthode que nous développons ici utilise les concepts de base de la théorie bayésienne décrits au deuxième chapitre et propose un critère de classification dérivant de la théorie des tests d'hypothèses. Nous présentons d'abord les principes de base de notre approche ainsi que l'algorithme que nous avons développé, d'un point de vue général. Dans tout le chapitre, nous considérons les variables comme étant des scalaires. Le cas multidimensionnel est présenté à la section 3.5. L'approche est identique : nous remplaçons la moyenne par un vecteur moyenne, et la variance par une matrice de variance-covariance. Nous présentons ensuite le modèle hiérarchique utilisé et le détail des calculs de chacune des lois dérivant de ce modèle. Le calcul des marginales introduites dans le critère de classification et leur approximation numérique seront présentés en détail. Enfin, nous voyons comment estimer de façon raisonnable les paramètres de nuisance de notre modèle, et comment une variation de ceux-ci permet à notre méthode de s'adapter au type de jeu de données considéré.

3.1. Principe et algorithme de notre approche

Comme dans toute classification, le but est de diviser un jeu de données en classes, à l'intérieur desquelles un certain critère d'homogénéité est vérifié. Contrairement aux méthodes usuelles que nous avons présentées dans les deux derniers chapitres, nous cherchons ici à classifier des individus dans le but d'obtenir une homogénéité au niveau de la moyenne et de la variance des variables. En d'autres termes, notre but est d'obtenir des groupes identifiables par une égalité (ou non) des moyennes des individus les composant, ainsi qu'une égalité (ou non) des variances. Concrètement, nous cherchons à établir quatre types de groupes différents : ceux à l'intérieur desquels les observations ont une espérance et une variance identiques, ceux pour lesquels l'espérance est identique mais la variance est différente, ceux ayant une moyenne différente et une variance identiques et enfin, les groupes pour lesquels ni la variance, ni l'espérance ne sont communes. Nous identifions donc chaque groupe par une homogénéité de la moyenne des individus, de la variance, ou bien des deux à la fois. Implicitement, le quatrième type de groupe s'établit comme un ensemble d'individus totalement hétérogènes, du point de vue de nos deux critères.

La méthode que nous avons développée est de type agglomératif, c'est-à-dire que nous classons initialement les N observations dans N classes, chacune d'entre elles contenant un seul élément (ce principe est défini à la section 1.2). Ces N groupes sont successivement couplés entre eux, réduisant à chaque couplage le nombre de groupes de 1. Dans les méthodes agglomératives usuelles, ce nombre décroît de N jusqu'à 1. Dans notre cas, le critère de classification que nous avons choisi se trouve être aussi un critère d'arrêt à chaque étape de l'algorithme. Ainsi, en fonction de ce dernier, le nombre de groupes final se fixe à un nombre g, qui diffère généralement de 1. De plus, afin d'éviter le problème, comme dans beaucoup

d'algorithmes étudiés, du nombre élevé de calculs à effectuer (dûs à un nombre de partitions possibles des données beaucoup trop important), nous choisissons d'effectuer une sorte de présélection en étudiant la matrice de similarité de nos observations, ou de nos classes d'observations. Nous testons alors notre critère de regroupement uniquement sur les individus les plus proches en terme de distance. L'algorithme que nous avons choisi se divise en trois étapes distinctes, successivement dépendantes les unes des autres.

i) Regroupement des observations en classes, à l'intérieur desquelles leur moyenne est " égale " (application d'un test d'égalité de moyennes).

ii) Subdivision des classes formées à l'étape précédente en fonction de la variance des observations. Nous obtenons donc, à l'intérieur de chacun des groupes créés à l'étape i), des sous-classes à l'intérieur desquelles les observations ont une même moyenne mais aussi une même variance. À la fin de cette deuxième étape, nous avons donc formé deux types de groupes. Pour ce faire, nous utilisons le test d'hypothèses suivant :

$$H_0 : \theta_1 = \theta_2, \sigma_1^2 = \sigma_2^2$$
$$\text{vs } H_1 : \theta_1 = \theta_2, \sigma_1^2 \neq \sigma_2^2,$$

où θ_i représente la moyenne de l'observation X_i et où σ_i^2 représente sa variance.

iii) Rassemblement des classes créés à l'étape ii) lorsque celles-ci n'appartiennent pas au même groupe (créé à l'étape i). Deux sous-classes satisfaisant cette condition sont deux groupes de deux moyennes différentes. Nous testons alors l'égalité ou non des variances de ces deux groupes. Les hypothèses

testées sont les suivantes :

$$H_0 : \theta_1 \neq \theta_2, \sigma_1^2 = \sigma_2^2$$

$$\text{vs } H_1 : \theta_1 \neq \theta_2, \sigma_1^2 \neq \sigma_2^2.$$

Avant de détailler chacune des étapes de cet algorithme, nous devons noter une distinction importante dans la notion d'égalité des moments dans chacune des étapes ci-dessus. À la première étape, la notion de moyenne égale se situe entre les observations de chacun des groupes. La notion de moyenne différente est donc inter-groupe. À la deuxième étape, lorsque nous parlons de variance égale, nous parlons des individus à l'intérieur de chaque groupe. La notion de variance différente est donc là aussi inter-groupes. Enfin, à la troisième étape, lorsque nous cherchons à regrouper les classes formées à la première étape, la notion de moyenne différente et de variance égale ou différente est entre les sous-groupes, et non entre les individus eux-même. Nous notons aussi que l'algorithme est valable pour les groupes de taille 1, puisque nous possédons une information *a priori* sur la variance et la moyenne de chaque observation.

Détaillons maintenant chacune des trois étapes, d'un point de vue algorithmique (les programmes peuvent être lus en annexe A).

- **Étape 1**
  1) Création de la matrice de similarité des données.
  2) Test de l'égalité des moyennes entre les deux individus les plus similaires. Si le test est positif, c'est à dire que nous acceptons l'hypothèse H_0, ces deux individus sont regroupés, nous recalculons la matrice de similarité des données (en utilisant la méthode centroïde) et nous recommençons le test sur les deux nouvelles observations les plus proches, et ainsi de suite jusqu'à ce que nous rejetions l'hypothèse d'égalité des moyennes. Nous supposons bien sûr que si nous rejetons l'hypothèse nulle pour

deux variables, dont la distance entre elles est d, nous rejetterons cette hypothèse pour tous les couples de variables séparés par une distance $d + \epsilon$ ($\epsilon > 0$). La partition que nous obtenons est donc formée d'un certain nombre de classes contenant deux éléments ainsi que des classes contenant un unique individu (classes initiales).

3) Chacune des classes formées est maintenant considérée comme un individu dont la valeur est la moyenne des individus la composant et dont le poids est égal au nombre d'individus (ici 2). Nous réitérons l'étape 2) (c'est-à-dire que nous regroupons les classes entre elles, par couple) jusqu'à ce que nous ne puissions plus former des groupes de deux classes ou de deux individus (ceci correspond alors à un rejet de l'hypothèse nulle).

- **Étape 2**

L'algorithme ci-dessous est appliqué à chacune des classes formées à l'étape 1. Nous explicitons ici la procédure pour une classe particulière que nous cherchons donc à subdiviser selon la variance des observations la composant.

1) Calcul de la variance S_0 du groupe.
2) Construction d'un vecteur jouant le même rôle que la matrice de similarité : la i^{e} composante de ce vecteur représente la variance (ou la trace de la matrice de variance-covariance, dans le cas multivarié) de la classe après avoir ôté la i^{e} observation. Nous pouvons donc savoir quelles sont les observations dont l'absence produit le plus grand écart de variance. Nous testons donc l'égalité des moyennes et des variances entre le groupe entier et le groupe sans la i^{e} observation. Si l'hypothèse nulle est rejetée (la différence de variance est donc significative), l'observation est mise de coté. Nous continuons alors ce processus avec l'observation suivante

jusqu'à ce que nous acceptions l'hypothèse d'égalité des variances. Nous obtenons donc à ce stade deux sous-groupes : le premier, à l'intérieur duquel les moyennes et les variances des observations sont égales et le deuxième que nous allons essayer de subdiviser de la même façon.

3) Réitération de 2) sur le deuxième sous-groupe obtenu.

4) Toute cette procédure est répétée jusqu'à ce qu'on ne puisse plus subdiviser les groupes, soit parce que la taille de ceux-ci est trop petite, soit parce que l'hypothèse d'égalité des variances est acceptée.

- **Étape 3**

  Nous cherchons ici à regrouper des sous-groupes crées à la première étape si ceux-ci ne font pas partie de la même classe. La procédure de regroupement est identique à celle de la première étape, en considérant une sous-classe comme un individu dont le poids est le nombre d'éléments la composant et dont la valeur est égale à la moyenne de ses éléments.

Exemple 3.1.1. *Nous traitons ici un petit exemple afin d'illustrer chacune des étapes de l'algorithme que nous avons présenté. Cet exemple est fictif et les valeurs numériques ainsi que les résultats des différents tests sont posés de façon arbitraire. Considérons 8 observations* $X_1, X_2, ..., X_8$.

- *Première étape : Création de groupes homogènes par la moyenne des observations. Cette étape est illustrée au tableau* **??**. *Nous obtenons donc trois classes :* $(X_1X_2X_3X_4)$, $(X_5X_6X_7)$ *et* (X_8).

TABLEAU 3.1.1. *Illustration de la première étape de l'algorithme*

ÉTAPE	*ÉVÉNEMENT*	*PARTITION*
1	*Construction de la matrice de similarité D*	$(X_1)(X_2)(X_3)...(X_8)$
2	*Éléments les plus proches :* X_1 *et* X_2 *Test égalité des moyennes :* H_0 *acceptée*	$(X_1X_2)(X_3)(X_4)...(X_8)$
3	*Reconstruction de la matrice D* *Éléments suivants les plus proches :* X_3 *et* X_4 *Test égalité des moyennes :* H_0 *acceptée*	$(X_1X_2)(X_3X_4)(X_5)...(X_8)$
4	*Reconstruction de la matrice D* *Éléments suivants les plus proches :* X_5 *et* X_6 *Test égalité des moyennes :* H_0 *acceptée*	$(X_1X_2)(X_3X_4)(X_5X_6)(X_7)(X_8)$
5	*Reconstruction de la matrice D* *Éléments suivants les plus proches :* X_7 *et* X_8 *Test égalité des moyennes :* H_0 *rejetée*	$(X_1X_2)(X_3X_4)(X_5X_6)(X_7)(X_8)$
6	*Reconstruction de la matrice D* *Éléments les plus proches :* (X_1X_2) *et* (X_3X_4) *Test égalité des moyennes :* H_0 *acceptée*	$(X_1X_2X_3X_4)(X_5X_6)(X_7)(X_8)$
7	*Reconstruction de la matrice D* *Éléments suivants les plus proches :* (X_5X_6) *et* X_7 *Test égalité des moyennes :* H_0 *acceptée*	$(X_1X_2X_3X_4)(X_5X_6X_7)(X_8)$
8	*Reconstruction de la matrice D* *Éléments les plus proches :* $(X_1X_2X_3X_4)$ *et* (X_8) *Test égalité des moyennes :* H_0 *rejetée*	$(X_1X_2X_3X_4)(X_5X_6X_7)(X_8)$

– *Deuxième étape : Division de ces groupes en sous-groupes de variance égale (voir tableau ??). Nous notons* $S_{[-i]}$ *la variance du groupe sans la* i^e *observation et* S_0 *la variance totale.*

TABLEAU 3.1.2. *Illustration de la deuxième étape de l'algorithme*

GROUPE	*ÉVÉNEMENT*	*PARTITION*
GROUPE 1 $(X_1X_2X_3X_4)$	*Calcul de la variance totale et de la variance en ôtant une observation* $S_0 = 5$, $S_{[-1]} = 1$, $S_{[-2]} = 2$, $S_{[-3]} = 3$, $S_{[-4]} = 4$.	$(X_1X_2X_3X_4)$
	Plus grand écart observé : X_1 *est ôté* *Test égalité des moyennes et des variances sur les groupes* $(X_2X_3X_4)$ *et* $(X_1X_2X_3X_4)$ *:* H_0 *rejetée*	$(X_2X_3X_4)(X_1)$
	Reconstruction du vecteur S *et de* S_0 *Plus grand écart observé :* X_2 *est ôté.* *Test sur les groupes* $(X_2X_3X_4)$ *et* (X_3X_4) *:* H_0 *rejetée*	$(X_3X_4)(X_1X_2)$
	Reconstruction du vecteur S *et de* S_0 *Plus grand écart observé :* X_3 *est ôté.* *Test sur les groupes* (X_2X_4) *et* (X_4) *:* H_0 *acceptée*	$(X_3X_4)(X_1X_2)$
	Tentative de subdivision du sous-groupe (X_1X_2) *:* *Échec*	
GROUPE 2 $(X_5X_6X_7)$	*Variance totale :* $S_0 = 5$ $S_{[-5]} = 4$, $S_{[-6]} = 3$, $S_{[-7]} = 2$	
	Plus grand écart observé : X_7 *ôté* *Test sur* $(X_5X_6X_7)$ *et* (X_5X_6) H_0 *rejetée*	$(X_5X_6)(X_7)$
	Reconstruction du vecteur S *et de* S_0 *Plus grand écart observé :* X_6 *ôté* *Test sur* (X_5X_6) *et* (X_5) H_0 *acceptée*	$(X_5X_6)(X_7)$

– *Troisième étape : Regroupement des sous-groupes pour obtenir des classes de moyennes différentes et de variances égales. Au terme de la deuxième*

étape, nous obtenons la partition suivante :

$$[(X_1X_2)(X_3X_4)][(X_5X_6)(X_7)][X_8]$$

Cette notation signifie que 3 groupes ont été crées : $[X_1X_2X_3X_4]$,$[X_5X_6X_7]$ *et* $[X_8]$. *À l'intérieur de ces groupes, les observations ont la même moyenne. Avec ces groupes, des sous-groupes ont été formés, symbolisés par des parenthèses. Tous les sous-groupe appartenant au même groupe ont donc la même moyenne, mais des variances différentes.*
Étant donné que nous cherchons à obtenir des classes dont la moyenne est différente, nous ne pouvons pas regrouper (X_1X_2) *et* (X_3X_4) *puisque ceux-ci ont la même moyenne (voir étape 1). Nous calculons donc la matrice de similarité au niveau de la variance des groupes notée* D. *L'élément* (i, j) *de cette matrice représente alors la distance entre la variance des sous-groupes* i *et* j. *Cette matrice ne tient bien sûr pas compte des groupes* (X_1X_2) *et* (X_3X_4), *ni de* (X_5X_6) *et* (X_7).

- *Éléments les plus proches en terme de variance :* (X_1X_2) *et* (X_5X_6).
- *Test d´égalité des variances :* H_0 *acceptée.*
- *Reconstruction de* D. *Éléments les plus proches en terme de variance :* (X_3X_3) *et* (X_7).
- *Test d´égalité des variances :* H_0 *rejetée.*
- *Fin du regroupement.*

Le résultat final est obtenu au tableau 3.1.3.

Enfin, selon nos observations, et selon le degré d'homogénéité que nous cherchons à obtenir, nous pouvons faire varier certains paramètres initiaux de notre

TABLEAU 3.1.3. *Illustration de la dernière étape de l'algorithme*

Moyenne égale, variance égale	X_1 *et* X_2 X_3 *et* X_4 X_5 *et* X_6
Moyenne égale, variance différente	(X_1X_2) *et* (X_3X_4) (X_5X_6) *et* (X_7)
Moyenne différente, variance égale	(X_1X_2) *et* (X_5X_6)

modèle (paramètres de nuisance) afin de " forcer " en quelque sorte la classification à être plus ou moins sévère du point de vue du critère de regroupement. La méthode que nous utilisons est décrite en détail à la section 3.3.

3.2. LE MODÈLE

Lors de chacun des tests, nous considérons deux couples de variables (X_1, S_1) et (X_2, S_2). Chacune de ces variables est identifiée par le centroïde et la dispersion du groupe qu'elle représente. Notons aussi n_1 et n_2 la taille respective des deux groupes, et donc les poids respectifs des deux variables. Le modèle général hiérarchique que nous utilisons est le suivant : nous supposons les X_i normaux, de moyenne θ_i et de variance σ_i^2/n_i. La moyenne θ_i suit elle aussi une loi normale, de moyenne θ_0 et de variance σ_i^2/n_0. La variance σ_i^2 suit une loi inverse-gamma de paramètres α et β. Enfin, la loi de S_i, représentant la somme des carrés des écarts par rapport à la moyenne du groupe auquel X_i est attaché, est une loi gamma de paramètres $(n_i - 1)/2$ et $1/(2\sigma_i^2)$. Ce modèle, que nous venons de présenter, correspond aux lois *a priori* de type g, développées par Zellner (1971, 1986). Ces lois sont construites de façon à simplifier les calculs le plus possible. Les paramètres n_0, θ_0, α et β ne sont pas aléatoires mais entièrement spécifiés et sont considérés

comme des paramètres de nuisance. Nous avons :

$$S_i = \sum_{k=1}^{n_i} (x_k - X_i)^2,$$

où x_k $(k = 1, ..., n_i)$ sont les éléments du groupe dont X_i est le centroïde. Le fait que S_i suive une loi gamma provient du fait que

$$\frac{S_i}{\sigma_i^2} \sim \chi^2(n_i - 1),$$

ce qui est équivalent à

$$\frac{S_i}{\sigma_i^2} \sim G\left(\frac{(n_i - 1)}{2}, \frac{1}{2}\right).$$

Par un simple changement de variable, nous retrouvons facilement les paramètres de la loi de S_i. Notons les quatre hypothèses suivantes :

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \theta, \sigma_1^2 = \sigma_2^2 = \sigma^2, \\ H_1 : \theta_1 \neq \theta_2, \sigma_1^2 \neq \sigma_2^2, \\ H_2 : \theta_1 \neq \theta_2, \sigma_1^2 = \sigma_2^2 = \sigma^2, \\ H_3 : \theta_1 = \theta_2 = \theta, \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

À ces quatre hypothèses, nous ajoutons deux hypothèses n'impliquant que l'espérance des observations. Dans ces cas là, nous notons :

$$\begin{cases} H_0' : \theta_1 = \theta_2 = \theta, \\ H_1' : \theta_1 \neq \theta_2. \end{cases}$$

Le modèle, sous chacune de ces six hypothèses est le suivant (i varie de 1 à 2) :

$$\text{sous } H_0 \begin{cases} X_i|\theta,\sigma^2 \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n_i}\right), \\ \theta|\sigma^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{n_0}\right), \\ \sigma^2 \sim IG(\alpha,\beta), \\ S_i|\sigma^2 \sim G\left(\frac{(n_i-1)}{2}, \frac{1}{2\sigma^2}\right), \end{cases}$$

$$\text{sous } H_1 \begin{cases} X_i|\theta_i,\sigma_i^2 \sim \mathcal{N}\left(\theta_i, \frac{\sigma_i^2}{n_i}\right), \\ \theta_i|\sigma_i^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma_i^2}{n_0}\right), \\ \sigma_i^2 \sim IG(\alpha,\beta), \\ S_i|\sigma_i^2 \sim G\left(\frac{(n_i-1)}{2}, \frac{1}{2\sigma_i^2}\right), \end{cases}$$

$$\text{sous } H_2 \begin{cases} X_i|\theta_i,\sigma^2 \sim \mathcal{N}\left(\theta_i, \frac{\sigma^2}{n_i}\right), \\ \theta_i|\sigma^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{n_0}\right), \\ \sigma^2 \sim IG(\alpha,\beta), \\ S_i|\sigma^2 \sim G\left(\frac{(n_i-1)}{2}, \frac{1}{2\sigma^2}\right), \end{cases}$$

$$\text{sous } H_3 \begin{cases} X_i|\theta,\sigma_i^2 \sim \mathcal{N}\left(\theta, \frac{\sigma_i^2}{n_i}\right), \\ \theta|\sigma_i^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma_i^2}{n_0}\right), \\ \sigma_i^2 \sim IG(\alpha,\beta), \\ S_i|\sigma^2 \sim G\left(\frac{(n_i-1)}{2}, \frac{1}{2\sigma_i^2}\right), \end{cases}$$

$$\text{sous } H_0' \begin{cases} X_i|\theta,\sigma^2 \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n_i}\right), \\ \theta|\sigma^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{n_0}\right), \\ \sigma^2 \sim IG(\alpha,\beta), \end{cases}$$

$$\text{sous } H_1' \begin{cases} X_i|\theta_i,\sigma^2 \sim \mathcal{N}\left(\theta_i, \frac{\sigma^2}{n_i}\right), \\ \theta_i|\sigma^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{n_0}\right), \\ \sigma^2 \sim IG(\alpha,\beta). \end{cases}$$

Rappelons maintenant les trois densités des trois lois principales que nous utilisons. De façon générale, si $X \sim \mathcal{N}(\theta,\sigma^2)$, alors X a pour densité :

$$f_X(x|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right].$$

De même, si $\sigma^2 \sim IG(\alpha,\beta)$ alors σ^2 a pour densité :

$$\pi(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\exp[-\beta/\sigma^2]}{(\sigma^2)^{\alpha+1}}.$$

Enfin, si $S \sim G(n,V)$, alors S a pour densité :

$$\pi(S|V) = \frac{S^{n-1}}{\Gamma(n)} \exp[-SV].$$

3.3. Calcul des marginales

Le critère de classification que nous utilisons se présente comme étant un rapport de marginales représentant une probabilité et donc, se comparant à 1. Les tests que nous effectuons sont regroupés en trois cas différents :

$$\begin{cases} \text{i) } H_0' : \theta_1 = \theta_2 = \theta \text{ versus } H_1' : \theta_1 \neq \theta_2, \\ \text{ii) } H_0 : \theta_1 = \theta_2 = \theta,\ \sigma_1^2 = \sigma_2^2 = \sigma^2 \text{ versus } H_3 : \theta_1 = \theta_2 = \theta,\ \sigma_1^2 \neq \sigma_2^2, \\ \text{iii) } H_2 : \theta_1 \neq \theta_2,\ \sigma_1^2 = \sigma_2^2 = \sigma^2 \text{ versus } H_1 : \theta_1 \neq \theta_2,\ \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

Dans les deux derniers cas, nous considérons donc un test sur la variance sachant que les moyennes sont égales ou non, ce qui correspond aux trois étapes de l'algorithme. Par contre, à la première étape, nous nous intéressons uniquement aux moyennes des observations et non aux variances. Nous effectuons donc le test i). À la deuxième étape, nous cherchons à diviser les groupes dont la moyenne est égales en sous-groupes à l'intérieur desquels la variance est homogène. Nous utilisons donc le test ii). Enfin, le test iii) est utilisé à la dernière étape de notre algorithme. Soit $m_i = m_i(x_1, x_2, S_1, S_2)$ la marginale des observations sous l'hypothèse H_i $(i = 0, .., 3)$. De même, nous définissons $m'_i = m'_i(x_1, x_2)$ comme étant la marginale de nos observations sous l'hypothèse H'_i $(i = 0, 1)$. Les probabilités *a posteriori* utilisées sont les suivantes :

$$\text{Cas i)} \quad \left.\begin{array}{l} p'_0 = \dfrac{m'_0}{m'_0 + m'_1} \\ p'_1 = \dfrac{m'_1}{m'_0 + m'_1} \end{array}\right\} \Rightarrow p'_0 + p'_1 = 1,$$

$$\text{Cas ii)} \quad \left.\begin{array}{l} p_0 = \dfrac{m_0}{m_0 + m_3} \\ p_3 = \dfrac{m_3}{m_0 + m_3} \end{array}\right\} \Rightarrow p_0 + p_3 = 1,$$

$$\text{Cas iii)} \quad \left.\begin{array}{l} p_1 = \dfrac{m_1}{m_1 + m_2} \\ p_2 = \dfrac{m_2}{m_1 + m_2} \end{array}\right\} \Rightarrow p_0 + p_3 = 1.$$

Dans le premier cas, p'_0 et p'_1 représentent les probabilités *a posteriori* des hypothèses H'_0 et H'_1. Dans les deux derniers cas, elles représentent les probabilités que les variances soient égales ou non, sachant que les moyennes sont égales ou non. Dans le cas i), nous comparons les probabilités *a posteriori* p'_0 et p'_1 à $1/2$.

Les régions critiques sont donc définies telles que

$$\begin{cases} H'_0 \text{ acceptée si } p'_0 > \frac{1}{2} (\Rightarrow H'_1 \text{ refusée}), \\ H'_1 \text{ acceptée si } p'_1 > \frac{1}{2} (\Rightarrow H'_0 \text{ refusée}). \end{cases}$$

Dans les deux autres cas, nous devons faire face à un problème inhérent à la première étape de l'algorithme : puisque les variables sont regroupées en classes à l'intérieur desquelles la moyenne est identique, il s'en suit que les observations sont assez concentrées autour de la moyenne de la classe. La variance est donc assez faible à l'intérieur de chaque groupe. La séparation des groupes en sous-groupes se doit donc d'être beaucoup plus sévère, c'est à dire que la division des classes doit être privilégiée. Nous verrons dans la section 3.5 comment une variation des paramètres de nuisance du modèle peut directement être lié au niveau de sévérité de la classification. Dans les deux derniers cas, nous avons donc :

$$\begin{cases} H_0 \text{ acceptée si } p_0 > c \ (\Rightarrow H_1 \text{ refusée}), \\ H_1 \text{ acceptée si } p_1 < c \ (\Rightarrow H_0 \text{ refusée}), \end{cases}$$

où $c \in [1/2, 1]$.

De même,

$$\begin{cases} H_2 \text{ acceptée si } p_2 > c \ (\Rightarrow H_3 \text{ refusée}), \\ H_3 \text{ acceptée si } p_3 < c \ (\Rightarrow H_2 \text{ refusée}). \end{cases}$$

Ici, la valeur de la cote limite, c, n'est pas réellement fixée. Nous la déterminons en fait en fonction du jeu de données et du nombre de sous-groupes que l'on s'attend à obtenir. De façon générale, une valeur de c de 0,9 est satisfaisante. Détaillons maintenant le calcul de chacune des six marginales, m'_0, m'_1, m_0, m_1, m_2 et m_3.

3.3.1. Marginale sous H_0' : $\theta_1 = \theta_2$

Nous cherchons à calculer $m_0' = m_0'(x_1, x_2)$. Le modèle utilisé étant un modèle hiérarchique, la marginale est donc égale à

$$m_0' = \int_0^{+\infty} \int_{-\infty}^{+\infty} \pi(x_1|\theta,\sigma^2)\pi(x_2|\theta,\sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)\, d\theta\, d\sigma^2.$$

Nous avons donc :

$$\begin{aligned} m_0' &= \frac{\sqrt{n_0 n_1 n_2}}{(2\pi)^{3/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{(\sigma^2)^{3/2}} \exp\left[\frac{-1}{2\sigma^2}(n_1(x_1-\theta)^2 + n_2(x_2-\theta)^2\right. \\ &\quad \left. + n_0(\theta-\theta_0)^2)\right] \frac{\exp[-\beta/\sigma^2]}{(\sigma^2)^{\alpha+1}}\, d\theta\, d\sigma^2, \\ &= \frac{\sqrt{n_0 n_1 n_2}}{(2\pi)^{3/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{(\sigma^2)^{3/2+\alpha+1}} \exp[-\beta/\sigma^2] \exp[-S^2/2\sigma^2] \\ &\quad \exp\left[\frac{-1}{2\sigma^2}(n_1+n_2)(\theta-\bar{X})^2 + n_0(\theta-\theta_0)^2\right] d\sigma^2\, d\theta, \end{aligned}$$

où nous définissons

$$\bar{X} = \frac{n_1 X_1 + n_2 X_2}{n_1 + n_2},$$

et

$$S^2 = n_1(X_1 - \bar{X})^2 + n_2(X_2 - \bar{X})^2.$$

Nous obtenons donc, en remarquant que les termes en θ sont proportionnels à une loi normale

$$\begin{aligned} m_0' &= \frac{\sqrt{n_0 n_1 n_2}}{(2\pi)^{3/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \frac{1}{(\sigma^2)^{3/2+\alpha+1}} \frac{(2\pi\sigma^2)^{1/2}}{\sqrt{n_0+n_1+n_2}} \\ &\quad \times \exp\left[\frac{-1}{2\sigma^2}\left\{\beta + \frac{S^2}{2} + \frac{n_0(n_1+n_2)}{2(n_0+n_1+n_2)}(\bar{X}-\theta_0)^2\right\}\right] d\sigma^2. \end{aligned}$$

Finalement, nous aboutissons à l'équation suivante, en remarquant que les termes en σ^2 sont proportionnels à une loi inverse gamma.

$$\begin{aligned} m_0' &= \frac{\sqrt{n_0 n_1 n_2}}{\sqrt{n_0+n_1+n_2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{2\pi} \\ &\quad \times \left[\beta + \frac{S^2}{2} + \frac{n_0(n_1+n_2)}{2(n_0+n_1+n_2)}(\bar{X}-\theta_0)^2\right]^{-(\alpha+1)}. \end{aligned}$$

3.3.2. Marginale sous H_1' : $\theta_1 \neq \theta_2$

Nous nous plaçons cette fois-ci dans le cas de l'hypothèse H_1', c'est-à-dire que nous supposons les deux moyennes différentes. Nous avons alors :

$$m_1' = \int_0^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \pi(x_1|\theta_1,\sigma^2)\pi(x_2|\theta_2,\sigma^2)\pi(\theta_1|\sigma^2)\pi(\theta_2|\sigma^2)\pi(\sigma^2)\, d\theta_1\, d\theta_2\, d\sigma^2.$$

Nous avons donc :

$$\begin{aligned} m_1' &= \frac{n_0\sqrt{n_1 n_2}}{(2\pi)^2}\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{(\sigma^2)^2} \\ &\quad \times \exp\left[\frac{-1}{2\sigma^2}(n_1(x_1-\theta_1)^2 + n_0(\theta_1-\theta_0)^2)\right] \\ &\quad \times \exp\left[\frac{-1}{2\sigma^2}(n_2(x_2-\theta_2)^2 + n_0(\theta_2-\theta_0)^2)\right] \frac{\exp[-\beta/\sigma^2]}{(\sigma^2)^{\alpha+1}}\, d\theta_1\, d\theta_2\, d\sigma^2. \end{aligned}$$

En combinant les termes en θ_1 et en θ_2 ensemble et en reconnaissant les densités de deux lois normales, nous avons alors :

$$\begin{aligned} m_1' &= \frac{n_0\sqrt{n_1 n_2}}{(2\pi)^2}\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} \frac{1}{(\sigma^2)^2}\frac{1}{(\sigma^2)^{\alpha+1}} \frac{2\pi\sigma^2}{\sqrt{(n_0+n_2)}\sqrt{(n_0+n_1)}} \\ &\quad \times \exp\left[\frac{-1}{2\sigma^2}\left(\beta + \frac{n_0 n_1}{2(n_0+n_1)}(x_1-\theta_0)^2 + \frac{n_0 n_2}{2(n_0+n_2)}(x_2-\theta_0)^2\right)\right] d\sigma^2. \end{aligned}$$

Nous obtenons finalement

$$m_1' = \frac{n_0\sqrt{n_1 n_2}}{\sqrt{(n_0+n_2)(n_0+n_1)}}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(2\pi)}$$

$$\times \left[\beta + \frac{n_0 n_1}{2(n_0+n_1)}(x_1-\theta_0)^2 + \frac{n_0 n_2}{2(n_0+n_2)}(x_2-\theta_0)^2\right]^{-(\alpha+1)}.$$

3.3.3. Marginale sous $H_0 : \theta_1 = \theta_2, \sigma_1^2 = \sigma_2^2$

Les marginales que nous calculons sous les quatre prochaines hypothèses sont aussi fonction des variances échantillonnales S_1 et S_2 des deux classes. Supposons les moyennes ainsi que les variances respectivement égales à θ et σ^2. La marginale se calcule alors de la façon suivante :

$$m_0 = \int_0^{+\infty}\int_{-\infty}^{+\infty} \pi(x_1|\theta,\sigma^2)\pi(x_2|\theta,\sigma^2)\pi(S_1|\sigma^2)\pi(S_2|\sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)\, d\theta\, d\sigma^2.$$

Il s'en suit donc que

$$\begin{aligned} m_0 = {} & \frac{\sqrt{n_0 n_1 n_2}}{(2\pi)^{3/2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)} \frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\ & \times \int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{(\sigma^2)^{3/2}} \exp\left[\frac{-1}{2\sigma^2}(n_1(x_1-\theta)^2 + n_2(x_2-\theta)^2 + n_0(\theta-\theta_0)^2)\right] \\ & \times \frac{\exp[-\beta/\sigma^2]}{(\sigma^2)^{\alpha+1}} \frac{\exp[-S_1/(2\sigma^2)]}{(2\sigma^2)^{(n_1-1)/2}} \frac{\exp[-S_2/(2\sigma^2)]}{(2\sigma^2)^{(n_2-1)/2}}\, d\theta\, d\sigma^2. \end{aligned}$$

En sachant que

$$\begin{aligned} & n_1(x_1-\theta)^2 + n_2(x_2-\theta)^2 + n_0(\theta-\theta_0)^2 \\ & \quad = (n_0+n_1+n_2)\left(\theta - \frac{(n_1x_1+n_2x_2+n_0\theta_0)}{n_0+n_1+n_2}\right)^2 \\ & + \frac{1}{n_0+n_1+n_2}\left(n_1n_2(x_1-x_2)^2 + n_0n_1(x_1-\theta_0)^2 + n_0n_2(x_2-\theta_0)^2\right), \end{aligned}$$

nous reconnaissons alors la densité d'une loi normale pour les termes en θ. Nous avons donc :

$$\begin{aligned} m_0 = {} & \frac{1}{2\pi} \frac{\sqrt{n_0 n_1 n_2}}{\sqrt{n_0+n_1+n_2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)} \frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \frac{1}{2^{(n_1+n_2)/2-1}} \\ & \int_0^{+\infty} \frac{1}{(\sigma^2)^{\alpha+1+(n_1+n_2)/2}} \exp\left[\beta + \frac{(S_1+S_2)}{2}\right. \end{aligned}$$

$$+\frac{\left(n_1n_2(x_1-x_2)^2+n_0n_1(x_1-\theta_0)^2+n_0n_2(x_2-\theta_0)^2\right)}{n_0+n_1+n_2}\Bigg]\, d\sigma^2.$$

En reconnaissant la densité d'une loi inverse gamma pour le paramètre σ^2, nous obtenons donc finalement :

$$\begin{aligned} m_0 \;=\; & \frac{1}{2\pi}\frac{\sqrt{n_0n_1n_2}}{(n_0+n_1+n_2)}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)}2^{\alpha+1} \\ & \times\Gamma(\alpha+(n_1+n_2)/2)\,[2\beta+S_1+S_2 \\ & +\frac{\left(n_1n_2(x_1-x_2)^2+n_0n_1(x_1-\theta_0)^2+n_0n_2(x_2-\theta_0)^2\right)}{n_0+n_1+n_2}\Bigg]^{-(\frac{(n_1+n_2)}{2}+\alpha)}. \end{aligned}$$

3.3.4. Marginale sous $H_1 : \theta_1 \neq \theta_2, \sigma_1^2 \neq \sigma_2^2$

Calculons maintenant la marginale m_1 sous l'hypothèse où les deux groupes sont totalement hétérogènes au niveau des variances et des moyennes. Nous obtenons donc :

$$\begin{aligned} m_1 \;=\; & \int_0^{+\infty}\int_0^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\pi(x_1|\theta_1,\sigma_1^2)\pi(x_2|\theta_2,\sigma_2^2)\pi(S_1|\sigma_1^2)\pi(S_2|\sigma_2^2) \\ & \times\pi(\theta_1|\sigma_1^2)\pi(\theta_2|\sigma_2^2)\pi(\sigma_1^2)\pi(\sigma_2^2)\,d\theta_1\,d\theta_2\,d\sigma_1^2\,d\sigma_2^2. \end{aligned}$$

Nous avons donc :

$$\begin{aligned} m_1 \;=\; & \frac{n_0\sqrt{n_1n_2}}{(2\pi)^2}\frac{\beta^{2\alpha}}{\Gamma^2(\alpha)}\frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\ & \times\int_0^{+\infty}\int_0^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{1}{\sigma_1^2}\exp\left[\frac{-1}{2\sigma_1^2}(n_1(x_1-\theta_1)^2+n_0(\theta_1-\theta_0)^2)\right] \\ & \times\frac{1}{\sigma_2^2}\exp\left[\frac{-1}{2\sigma_2^2}(n_2(x_2-\theta_2)^2+n_0(\theta_2-\theta_0)^2)\right] \\ & \times\frac{\exp[-\beta/\sigma_1^2]}{(\sigma_1^2)^{\alpha+1}}\frac{\exp[-\beta/\sigma_2^2]}{(\sigma_2^2)^{\alpha+1}} \\ & \times\frac{\exp[-S_1/(2\sigma_1^2)]}{(2\sigma_1^2)^{(n_1-1)/2}}\frac{\exp[-S_2/(2\sigma_2^2)]}{(2\sigma_2^2)^{(n_2-1)/2}}\,d\theta_1\,d\theta_2\,d\sigma_1^2\,d\sigma_2^2. \end{aligned}$$

En procédant de la même façon que pour les intégrales précédentes, c'est-à-dire en reconnaissant successivement les densités de lois normales pour les termes θ_1 et θ_2 puis des lois inverses gamma pour les termes en σ_1^2 et σ_2^2, nous obtenons finalement :

$$\begin{aligned} m_1 &= \frac{n_0\sqrt{n_1 n_2}}{\sqrt{(n_0+n_1)(n_0+n_2)}}\frac{\beta^{2\alpha}}{\Gamma^2(\alpha)} \\ &\times \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)}\frac{2^{(2\alpha+(n_1+n_2)/2)}}{2\pi} \\ &\times \frac{\Gamma(\alpha+n_1/2)}{\left(2\beta+S_1+\dfrac{n_0 n_1}{n_0+n_1}(x_1-\theta_0)^2\right)^{(\alpha+n_1/2)}} \\ &\times \frac{\Gamma(\alpha+n_2/2)}{\left(2\beta+S_2+\dfrac{n_0 n_2}{n_0+n_2}(x_2-\theta_0)^2\right)^{(\alpha+n_2/2)}}. \end{aligned}$$

3.3.5. Marginale sous $H_2 : \theta_1 \neq \theta_2, \sigma_1^2 = \sigma_2^2$

Nous considérons maintenant le cas des moyennes différentes mais des variances égales à σ^2. La marginales s'écrit donc sous cette forme :

$$\begin{aligned} m_2 &= \int_0^{+\infty}\int_0^{+\infty}\int_{-\infty}^{+\infty} \pi(x_1|\theta_1,\sigma^2)\pi(x_2|\theta_2,\sigma^2)\pi(S_1|\sigma^2)\pi(S_2|\sigma^2) \\ &\quad \times\pi(\theta_1|\sigma^2)\pi(\theta_2|\sigma^2)\pi(\sigma^2)\,d\theta_1\,d\theta_2\,d\sigma^2. \end{aligned}$$

Toujours en procédant de la même façon, nous obtenons :

$$\begin{aligned} m_2 &= \frac{n_0\sqrt{n_1 n_2}}{\sqrt{(n_0+n_1)(n_0+n_2)}}\frac{2^{\alpha+(n_1+n_2)/2}}{2\pi}\frac{\beta^{\alpha}}{\Gamma(\alpha)} \\ &\times \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \end{aligned}$$

$$\times \frac{\Gamma(\alpha + (n_1+n_2)/2)}{\left[S_1 + S_2 + 2\beta + \frac{n_1 n_0}{n_1+n_0}(x_1-\theta_0)^2 + \frac{n_2 n_0}{n_2+n_0}(x_2-\theta_0)^2\right]^{\alpha+(n_1+n_2)/2}}.$$

3.3.6. Marginale sous H_3 : $\theta_1 = \theta_2, \sigma_1^2 \neq \sigma_2^2$

Cette dernière densité, dans le cas où les moyennes sont égales à θ et où les variances sont différentes n'est pas calculable analytiquement. Nous recourons donc à une approximation numérique. La forme générale de cette densité est :

$$\begin{aligned} m_3 &= \int_0^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \pi(x_1|\theta,\sigma_1^2)\pi(x_2|\theta,\sigma_2^2)\pi(S_1|\sigma_1^2)\pi(S_2|\sigma_2^2) \\ &\quad \times\pi(\theta|\sigma_1^2)\pi(\theta|\sigma_2^2)\pi(\sigma_1^2)\pi(\sigma_2^2)\, d\theta\, d\sigma_1^2\, d\sigma_2^2. \end{aligned}$$

La densité de (x_1, x_2) conditionnellement aux trois paramètres θ, σ_1^2, σ_2^2 peut s'écrire comme étant le produit suivant (du fait de l'indépendance de nos deux observations) :

$$\pi(x_1, x_2|\theta, \sigma_1^2, \sigma_2^2) = \pi(x_1|\theta,\sigma_1^2)\pi(x_2|\theta,\sigma_2^2).$$

Pour calculer ce produit, nous utilisons la relation suivante :

$$\frac{n_1}{\sigma_1^2}(x_1-\theta)^2 + \frac{n_2}{\sigma_2^2}(x_2-\theta)^2 =$$
$$\frac{(n_1\sigma_2^2+n_2\sigma_1^2)}{\sigma_1^2\sigma_2^2}\left(\theta - \frac{n_1\sigma_2^2x_1 + n_2\sigma_1^2x_2}{n_1\sigma_2^2+n_2\sigma_1^2}\right)^2 + \frac{n_1n_2}{n_1\sigma_2^2+n_2\sigma_1^2}(x_1-x_2)^2.$$

De plus, la loi de θ conditionnellement aux deux paramètres σ_1^2 et σ_2^2 est la suivante :

$$\theta|\sigma_1^2\sigma_2^2 \sim \mathcal{N}\left(\theta_0, \frac{\sigma_1^2\sigma_2^2}{n_0(n_1\sigma_2^2+n_2\sigma_1^2)}\right).$$

Nous avons donc :

$$m_3 = \int_0^{+\infty}\int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{\sqrt{n_1n_2}}{2\pi}\exp\left[\frac{-(n_1\sigma_2^2+n_2\sigma_1^2)}{2\sigma_1^2\sigma_2^2}\left(\theta - \frac{n_1\sigma_2^2x_1+n_2\sigma_1^2x_2}{n_1\sigma_2^2+n_2\sigma_1^2}\right)^2\right]$$

$$\times \exp\left[\frac{n_1 n_2}{2(n_1\sigma_2^2 + n_2\sigma_1^2)}(x_1 - x_2)^2\right] \frac{1}{\sigma_1^2\sigma_2^2}$$
$$\times \frac{\sqrt{n_0}}{\sqrt{2\pi}} \frac{\sqrt{n_1\sigma_2^2 + n_2\sigma_1^2}}{\sigma_1^2\sigma_2^2} \exp\left[\frac{-n_0(n_1\sigma_2^2 + n_2\sigma_1^2)}{2\sigma_1^2\sigma_2^2}(\theta_1 - \theta_2)^2\right]$$
$$\times \frac{\beta^{2\alpha}}{\Gamma^2(\alpha)} \frac{\exp[-\beta/\sigma_1^2]}{(\sigma_1^2)^{\alpha+1}} \frac{\exp[-\beta/\sigma_2^2]}{(\sigma_2^2)^{\alpha+1}} \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)} \frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \frac{1}{2^{(n_1+n_2)/2+1}}$$
$$\times \frac{\exp[-S_1/(2\sigma_1^2)]}{(2\sigma_1^2)^{(n_1-1)/2}} \frac{\exp[-S_2/(2\sigma_2^2)]}{(2\sigma_2^2)^{(n_2-1)/2}} \, d\theta \, d\sigma_1^2 \, d\sigma_2^2.$$

Lorsque nous calculons le terme en θ dans les deux exponentielles, nous obtenons une exponentielle dont l'exposant est :

$$\frac{n_1\sigma_2^2 + n_2\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left[(1+n_0)\left(\theta - \frac{A + n_0\theta_0}{1+n_0}\right)^2 + \frac{n_0}{1+n_0}(\theta_0 - A)^2\right],$$

$$\text{où } A = \frac{n_1\sigma_2^2 x_1 + n_2\sigma_1^2 x_2}{n_1\sigma_2^2 + n_2\sigma_1^2}.$$

En reconnaissant alors que les termes en θ sont proportionnels à une loi normale, nous avons alors

$$\begin{aligned} m_3 \;=\; & \frac{\sqrt{n_0 n_1 n_2}}{2\pi} \frac{\beta^{2\alpha}}{\Gamma^2(\alpha)} \frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)} \frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\ & \times \frac{1}{\sqrt{1+n_0}} \exp\left[\frac{-n_1 n_2}{2(n_1\sigma_2^2 + n_2\sigma_1^2)}(x_1 - x_2)^2\right] \\ & \times \int_0^{+\infty}\int_0^{+\infty} \frac{1}{(\sigma_1^2)^{\alpha+1+n_1/2}} \frac{1}{(\sigma_2^2)^{\alpha+1+n_2/2}} \\ & \times \exp\left[\frac{-1}{2\sigma_1^2}(2\beta + S_1)\right] \exp\left[\frac{-1}{2\sigma_2^2}(2\beta + S_2)\right] \\ & \times \exp\left[\frac{-(n_1\sigma_2^2 + n_2\sigma_1^2)}{2\sigma_1^2\sigma_2^2} \frac{n_0}{1+n_0}(\theta_0 - A)^2\right] d\sigma_1^2 \, d\sigma_2^2. \end{aligned}$$

Effectuons alors le changement de variable suivant :

$$\sigma_2^2 = V\sigma_1^2 \Rightarrow d\sigma_2^2 = \sigma_1^2 dV.$$

Alors,

$$
\begin{aligned}
m_3 &= \frac{\sqrt{n_0 n_1 n_2}}{2\pi}\frac{\beta^{2\alpha}}{\Gamma^2(\alpha)}\frac{1}{\sqrt{1+n_0}}\frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\
&\quad \times \int_0^{+\infty}\int_0^{+\infty} \frac{1}{(\sigma_1^2)^{2\alpha++1+(n_1+n_2)/2}}\frac{1}{V^{\alpha+1+n_2/2}} \\
&\quad \times \exp\left[\frac{-1}{2\sigma_1^2 V}\left(2V\beta + S_1 V + 2\beta + S_2\right.\right. \\
&\quad + \frac{n_0(n_1V+n_2)}{1+n_0}\left(\theta_0 - \frac{n_1 V x_1 + n_2 x_2}{n_1 V + n_2}\right)^2 \\
&\quad \left.\left. + \frac{n_1 n_2 V}{n_1 V + n_2}(x_1 - x_2)^2\right)\right] d\sigma_1^2\, dV.
\end{aligned}
$$

Nous remarquons alors que le terme en σ_1^2 est proportionnel à une loi inverse gamma. Nous avons donc

$$
\begin{aligned}
m_3 &= \frac{\sqrt{n_0 n_1 n_2}}{2\pi}\frac{\beta^{2\alpha}}{\Gamma^2(\alpha)}\frac{1}{\sqrt{1+n_0}}\frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\
&\quad \times \Gamma(2\alpha + (n_1+n_2)/2)2^{2\alpha+(n_1+n_2)/2}\int_0^{+\infty}\frac{1}{V^{\alpha+1+n_2/2}} \\
&\quad \times \left[2\beta\frac{(V+1)}{V} + S_1 + \frac{S_2}{V}\right. \\
&\quad \left. + \frac{(n_1V+n_2)}{V}\frac{n_0}{1+n_0}\left(\theta_0 - \frac{n_1 V x_1 + n_2 x_2}{n_1 V + n_2}\right)^2\right]^{-(2\alpha+\frac{n_1+n_2}{2})} dV.
\end{aligned}
\tag{3.3.1}
$$

Cette intégrale ne peut se calculer analytiquement. Nous devons donc recourir à une approximation numérique du type Monte-Carlo.

3.3.7. Méthode de Monte Carlo : approximation de la densité m_3

La méthode de Monte Carlo est une technique classique de calcul bayésien à la fois simple et efficace qui permet d'approcher de façon numérique des intégrales non calculables analytiquement. Le principe de la méthode est le suivant (voir

Robert, 1992). Nous cherchons à calculer une intégrale de la forme :

$$\int_{\Theta} g(\theta) f(x|\theta) \pi(\theta)\, d\theta,$$

que l'on peut aussi écrire

$$\int_{\Theta} \frac{g(\theta) f(x|\theta) \pi(\theta)}{h(\theta)} h(\theta)\, d\theta, \tag{3.3.2}$$

où h est une densité de probabilité sur Θ vérifiant les trois propriétés suivantes :

i) le support de $f(x|\theta)\pi(\theta)$ est inclus dans le support de h,

ii) $\lim_{|\theta| \to \infty} \frac{f(x|\theta)\pi(\theta)}{h(|\theta|)} = 0$,

iii) il doit être relativement facile, d'un point de vue numérique, de générer θ à partir de h.

Nous appelons alors méthode de Monte Carlo de fonction d'importance h l'algorithme suivant :

1) Générer $\theta_1, ..., \theta_m$ selon la densité h,

2) l'intégrale (3.3.2) est approchée par :

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) w(\theta_i),$$

où $w(\theta_i) = \frac{f(x|\theta_i)\pi(\theta_i)}{h(\theta_i)}$.

La loi forte des grands nombres assure, sous quelques conditions de régularité, comme le fait que l'intégrale à calculer soit finie, que cette approximation converge vers l'équation (3.3.2). Dans notre cas, en reprenant l'intégrale m_3 calculée à la section 3.3.6, nous pouvons poser

$$m_3 = \int_0^{\infty} f(V)\, dV,$$

où f est la fonction que nous cherchons à intégrer à l'équation (3.3.1). Si nous choisissons V suivant une loi de Fisher de paramètres $\nu = 2\alpha + n_1$ et $\rho = 2\alpha + n_2$

et posons h la fonction de densité de cette loi, nous avons alors

$$\begin{aligned} h(V) &= \frac{\Gamma((2\alpha+n_1+n_2)/2)(\alpha+n_1)^{(\alpha+n_1)/2}(\alpha+n_2)^{(\alpha+n_2)/2}}{\Gamma((\alpha+n_1)/2)\Gamma((\alpha+n_2)/2)} \\ &\quad \times \frac{V^{(\alpha+n_1-2)/2}}{(\alpha+n_1+V(\alpha+n_2))^{(2\alpha+n_1+n_2)/2}}. \end{aligned} \tag{3.3.3}$$

Nous pouvons alors écrire

$$\begin{aligned} m_3 &= \int_0^\infty \frac{f(V)}{h(V)} h(V)\, dV, \\ &= \int_0^\infty w(V) h(V)\, dV, \\ &\simeq \frac{1}{m}\sum_{i=1}^{m} w(V_i). \end{aligned}$$

Nous avons donc

$$\begin{aligned} m_3 &= \frac{\sqrt{n_0 n_1 n_2}}{1+n_0}\frac{\beta^{2\alpha}}{\Gamma^2(\alpha)}\frac{1}{2\pi}\frac{S_1^{(n_1-3)/2}}{\Gamma((n_1-1)/2)}\frac{S_2^{(n_2-3)/2}}{\Gamma((n_2-1)/2)} \\ &\quad \times \frac{\Gamma(2\alpha+(n_1+n_2)/2)2^{2\alpha+(n_1+n_2)/2}\Gamma((\alpha+n_1)/2)\Gamma((\alpha+n_2)/2)}{\Gamma((2\alpha+n_1+n_2)/2)(\alpha+n_1)^{(\alpha+n_1)/2}(\alpha+n_2)^{(\alpha+n_2)/2}} \\ &\quad \times \sum_{i=1}^{1000}\left[2\beta\frac{(V_i+1)}{V_i} + S_1 + \frac{S_2}{V_i} + \right. \\ &\quad \left. \frac{n_1V_i+n_2}{V_i}\frac{n_0}{1+n_0}\left(\theta_0 - \frac{n_1V_ix_1+n_2x_2}{n_1V_i+n_2}\right)^2\right]^{-(2\alpha+(n_1+n_2)/2)} \\ &\quad \times \frac{(\alpha+n_1+V_i(\alpha+n_2))^{(2\alpha+n_1+n_2)/2}}{V_i^{3\alpha/2+(n_1+n_2)/2}}, \end{aligned}$$

où les V_i sont 1000 variables aléatoires générées suivant la loi de Fisher décrite à l'équation (3.3.3).

3.4. Estimation des paramètres de nuisance

Comme nous l'avons vu, le modèle que nous utilisons est un modèle hiérarchique dont les quatre paramètres de nuisance sont α, β, n_0 et θ_0. Dans un tel modèle, ces paramètres sont totalement spécifiés. Dans notre cas, afin de ne pas faire un choix qui serait purement arbitraire, nous essayons de fixer ces paramètres en fonction du jeu de données disponible. Les deux paramètres θ_0 et β peuvent être calculés par la méthode des moments. Ce n'est malheureusement pas le cas de α et n_0 dont nous étudierons plus tard l'influence à l'aide de tables de valeurs que nous avons construites. En ce qui concerne θ_0, nous avons les relations suivantes :

$$\begin{cases} \mathbb{E}[X_i|\theta_i] = \theta_i, \\ \mathbb{E}[\theta_i] = \theta_0. \end{cases}$$

Nous pouvons donc facilement approximer θ_0 par la moyenne des observations, c'est-à-dire :

$$\theta_0 = \frac{n_1 X_1 + n_2 X_2}{n_1 + n_2}.$$

D'autre part, nous avons

$$\begin{aligned} \mathbb{E}[S_i|\sigma_i^2] &= (n_i - 1)\sigma_i^2, \\ \text{et } \mathbb{E}[\sigma_i^2] &= \frac{\beta}{(\alpha - 1)} \text{ (espérance d'une loi } IG(\alpha, \beta)). \end{aligned}$$

Posons $Z_i = S_i/(n_i - 1)$. Nous avons alors

$$\mathbb{E}[Z_i] = \mathbb{E}[\sigma_i^2] = \frac{\beta}{(\alpha - 1)}.$$

Par la méthode des moments, nous devons résoudre

$$\frac{(Z_1 + Z_2)}{2} = \frac{\beta}{(\alpha - 1)}.$$

Nous avons donc finalement

$$\beta = (\alpha - 1)\frac{(Z_1 + Z_2)}{2}.$$

La méthode que nous avons choisie pour définir les valeurs de α et n_0 est la suivante : posons $S_1 = S_2 + \Delta$. Pour une différence de variance Δ fixée, nous calculons la valeur de m_0 en posant $x_1 = x_2 = 0$ (cas des moyennes égales) pour plusieurs valeurs de α et n_0. Pour une différence $\Delta = 0$ par exemple, nous choisissons alors la combinaison (n_0, α) pour laquelle la valeur de m_0 est la plus élevée. Nous verrons des exemples au dernier chapitre. Cette méthode, bien sûr, ne prétend pas fournir des estimations très précises des paramètres mais elle permet plutôt de ne pas assigner des valeurs arbitraires.

3.5. Approche multivariée

Jusqu'à présent, nous avons développé une approche bayésienne de classification dans le cas unidimensionnel. Nous allons maintenant étendre l'algorithme et le modèle présentés à un cadre multivarié. En réalité, l'approche reste identique et les différentes étapes de l'algorithme restent inchangées. Les principales modifications résident donc dans la définition du modèle et par conséquent, dans le calcul des nouvelles densités. Soient $\underset{\sim}{X}_1, \underset{\sim}{X}_2$ les deux observations, éléments de $\mathbb{R}^p$. Le modèle général est le suivant (i=1,2) :

$$\begin{cases} \underset{\sim}{X}_i | \underset{\sim}{\theta}_i, \Sigma_i \sim \mathcal{N}_p\left(\underset{\sim}{\theta}_i, \frac{\Sigma_i}{n_i}\right), \\ \underset{\sim}{\theta}_i | \Sigma_i \sim \mathcal{N}_p\left(\underset{\sim}{\theta}_0, \frac{\Sigma_i}{n_0}\right), \\ \Sigma_i \sim IW(\alpha, B), \\ \mathbb{S}_i | \Sigma_i \sim W(n_i - 1, \Sigma_i), \end{cases}$$

où $\mathbb{S}_i$ est définie dans ce cas par

$$\mathbb{S}_i = \sum_{k=1}^{n_i} (\underset{\sim}{x}_k - \underset{\sim}{X}_i)(\underset{\sim}{x}_k - \underset{\sim}{X}_i)',$$

où $\underset{\sim}{x}_k$ $(k = 1, ..., n_i)$ sont les éléments du groupe dont $\underset{\sim}{X}_i$ est le centroïde. Les modèles sous chacune des six hypothèses sont donc identiques à ceux décrits à la section 3.2, en remplaçant les normales univariées par des normales multivariées, les lois inverse gamma par des lois inverse Wishart et les lois gamma par des lois Wishart. Les densités de chacune de ces lois sont données à la section 2.2. Le calcul de chacune des six marginales, sous leur hypothèse respective est identique au cas univarié. Nous présentons ici les six résultats :

$$\begin{aligned} m_0' &= \frac{1}{(2\pi)^p}\left(\frac{n_0 n_1 n_2}{n_0+n_1+n_2}\right)^{p/2} \frac{|B|^{\alpha/2}}{\Gamma_p(\alpha/2)} \frac{\Gamma_p(\alpha/2+1)}{2^{\alpha p/2}} \\ &\quad \times \left| B + \frac{n_1 n_2}{n_0+n_1+n_2}(\underset{\sim}{x}_1 - \underset{\sim}{x}_2)(\underset{\sim}{x}_1 - \underset{\sim}{x}_2)' + \frac{n_0 n_1}{n_0+n_1+n_2}(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)' \right. \\ &\quad \left. + \frac{n_0 n_2}{n_0+n_1+n_2}(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)' \right|^{-(1+\alpha/2)}. \end{aligned}$$

$$\begin{aligned} m_1' &= \frac{n_0^p}{(2\pi)^p}\left(\frac{n_1 n_2}{(n_0+n_1)(n_0+n_2)}\right)^{p/2} \frac{|B|^{\alpha/2}}{\Gamma_p(\alpha/2)} \frac{\Gamma_p((\alpha+1)/2)}{2^{\alpha p/2}} \\ &\quad \times \left| B + \frac{n_0 n_1}{n_0+n_1}(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)' \right. \\ &\quad \left. + \frac{n_0 n_2}{n_0+n_2}(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)' \right|^{-(\alpha+1)/2}. \end{aligned}$$

$$\begin{aligned} m_0 &= \frac{1}{(2\pi)^p}\left(\frac{n_0 n_1 n_2}{n_0+n_1+n_2}\right)^{p/2} \frac{|B|^{\alpha/2}}{\Gamma_p(\alpha/2)} 2^{p(n_1+n_2+\alpha)} \\ &\quad \times \frac{|\mathbb{S}_1|^{(n_1-p-2)/2}}{\Gamma_p((n_1-1)/2)} \frac{|\mathbb{S}_2|^{(n_2-p-2)/2}}{\Gamma_p((n_2-1)/2)} \frac{\Gamma_p((n_1+n_2+\alpha)/2)}{2^{p(n_1+n_2+\alpha-2)/2}} \end{aligned}$$

$$\times \left| B + \mathbb{S}_1 + \mathbb{S}_2 + \frac{n_1 n_2}{n_0 + n_1 + n_2} (\underset{\sim}{x}_1 - \underset{\sim}{x}_2)(\underset{\sim}{x}_1 - \underset{\sim}{x}_2)' \right.$$
$$+ \frac{n_0 n_1}{n_0 + n_1 + n_2} (\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)'$$
$$\left. + \frac{n_0 n_2}{n_0 + n_1 + n_2} (\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)' \right|^{-(n_1+n_2+\alpha)/2} .$$

$$m_1 = \frac{n_0^p}{(2\pi)^p} \left(\frac{n_1 n_2}{(n_0 + n_1)(n_0 + n_2)} \right)^{p/2} \frac{|B|^\alpha}{\Gamma_p^2(\alpha/2)} 2^{p(n_1+n_2-2+\alpha)}$$
$$\times \frac{|\mathbb{S}_1|^{(n_1-p-2)/2}}{\Gamma_p((n_1-1)/2)} \frac{|\mathbb{S}_2|^{(n_2-p-2)/2}}{\Gamma_p((n_2-1)/2)} \frac{\Gamma_p((n_1-1+\alpha)/2)}{2^{p(n_1+n_2-2)/2}} \Gamma_p((n_2-1+\alpha)/2)$$
$$\times \left| B + \mathbb{S}_1 + \frac{n_0 n_1}{n_0 + n_1} (\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)' \right|^{-(n_1-1+\alpha)/2}$$
$$\times \left| B + \mathbb{S}_2 + \frac{n_0 n_2}{n_0 + n_2} (\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)' \right|^{-(n_2-1+\alpha)/2} .$$

$$m_2 = \frac{n_0^p}{(2\pi)^p} \left(\frac{n_1 n_2}{(n_0 + n_1)(n_0 + n_2)} \right)^{p/2} \frac{|B|^{\alpha/2}}{\Gamma_p(\alpha/2)} \Gamma_p((n_1+n_2+\alpha)/2)$$
$$\times \frac{|\mathbb{S}_1|^{(n_1-p-2)/2}}{\Gamma_p((n_1-1)/2)} \frac{|\mathbb{S}_2|^{(n_2-p-2)/2}}{\Gamma_p((n_2-1)/2)} 2^{p(n_1+n_2+\alpha)/2+1}$$
$$\times \left| B + \mathbb{S}_1 + \mathbb{S}_2 + \frac{n_0 n_1}{n_0 + n_1} (\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_1 - \underset{\sim}{\theta}_0)' \right.$$
$$\left. + \frac{n_0 n_2}{n_0 + n_2} (\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)(\underset{\sim}{x}_2 - \underset{\sim}{\theta}_0)' \right|^{-(n_1+n_2+\alpha)/2} .$$

$$m_3 = \frac{(n_0 n_1 n_2)^p}{(2\pi)^{p/2}} \frac{|B|^\alpha}{\Gamma_p^2(\alpha/2)} \frac{2^{p(\alpha+p+(n_1+n_2)/2)}}{(1+n_0)^{p/2}}$$
$$\times \frac{|\mathbb{S}_1|^{(n_1-p-2)/2}}{\Gamma_p((n_1-1)/2)} \frac{|\mathbb{S}_2|^{(n_2-p-2)/2}}{\Gamma_p((n_2-1)/2)}$$
$$\times \Gamma_p \left(\frac{2\alpha + p + n_1 + n_2 - 1}{2} \right) \int_0^{+\infty} v^{-(\alpha+p+n_2+2)/2}$$

$$\left| B(1+1/v) + S_1 + \frac{S_2}{v} + n_1 n_2 (n_1 v + n_2)^{-1} (\underset{\sim}{x_1} - \underset{\sim}{x_2})(\underset{\sim}{x_1} - \underset{\sim}{x_2})' \right.$$
$$\left. + \frac{n_0}{v(1+n_0)} (n_1 v + n_2)(\underset{\sim}{\theta_0} - A)(\underset{\sim}{\theta_0} - A)' \right|^{-(2\alpha+p+n_1+n_2-1)/2} dv,$$

où $A = (n_1 v + n_2)^{-1}(n_1 v \underset{\sim}{x_1} + n_2 \underset{\sim}{x_2})$ et où v est un scalaire (nous avons fait l'hypothèse que $\Sigma_2 = v\Sigma_1$). Cette dernière intégrale, comme précédemment, n'est pas calculable analytiquement. Nous utilisons donc l'approximation de Monte Carlo. Nous choisissons cette fois-ci $v \sim F(p(2\alpha+p+n_1+n_2-1)+2, \alpha(2p+1)+pn_1+3)$. Posons g la densité correspondant à cette loi de Fisher. Posons aussi $\nu = p(2\alpha+p+n_1+n_2-1)+2$ et $\rho = \alpha(2p+1)+pn_1+3$. Nous avons alors

$$h(v) = \frac{\Gamma((\nu+\rho)/2)\nu^{\nu/2}\rho^{\rho/2}}{\Gamma(\nu/2)\Gamma(\rho/2)} \frac{v^{(\nu-2)/2}}{(\nu+\rho x)^{(\rho+\nu)/2}}. \tag{3.5.1}$$

Nous obtenons donc

$$m_3 = \frac{1}{m} \sum_{i=1}^{m} w(v_i),$$

où les v_i sont m variables générées selon la loi gamma définie à l'équation (3.5.1) et où w est définie de la façon suivante :

$$w(V) = \frac{f(V)}{h(V)},$$

où $f(V)$ est la fonction intégrée lors du calcul de m_3.

L'estimation des paramètres de nuisance pourrait se faire de façon identique au cas unidimensionnel, c'est-à-dire par la méthode des moments. Dans notre cas, afin de simplifier les calculs, nous choisissons α et n_0 à l'aide de tables et posons arbitrairement $B = I_p$, où I_p représente la matrice identité $p \times p$. Le paramètre θ_0, quant à lui est estimé très facilement par la méthode des moments, tel que

$$\underset{\sim}{\theta_n} = \frac{n_1 \underset{\sim}{X_1} + n_2 \underset{\sim}{X_2}}{n_1 + n_2}.$$

Nous verrons, dans le prochain chapitre, une application de cette méthode à deux jeux de données, univarié et multivarié. Nous comparerons ensuite les résultats obtenus avec les méthodes classiques étudiées au premier chapitre. cp -r labbea/memoire/latex

Chapitre 4

COMPARAISON DES DIFFÉRENTS ALGORITHMES

Ce chapitre a pour but d'appliquer et de comparer les méthodes fréquentistes et bayésiennes que nous avons présentées dans les chapitres précédents.

4.1. Présentation des données

Nous utilisons un jeu de données fréquemment utilisé en classification : les iris de Fisher. Ce jeu de données représente la mesure de quatre paramètres de 150 spécimens d'iris. Ces paramètres sont la longueur des pétales, la largeur des pétales, la longueur des sépales et la largeur des sépales. Trois espèces d'iris sont étudiées : Iris Setosa, Iris Versicolor et Iris Virginica. Le jeu de données contient donc 50 observations de chaque espèce. Il a donc l'avantage d'avoir une classification pré-établie en trois groupes, ce qui nous permet de comparer nos résultats et ceux des différentes méthodes que nous testons, à la classification réelle. Soient $X_1, X_2, ..., X_n$ $(n = 150)$ les données des trois types d'iris telles que $X_i \in \mathbb{R}^p$ $(p = 4)$. La répartition des groupes est alors la suivante :

$$\begin{cases} X_i \in GR1 \text{ si } i \in \{1, ...50\}, \\ X_i \in GR2 \text{ si } i \in \{51, ...100\}, \\ X_i \in GR3 \text{ si } i \in \{101, ...150\}, \end{cases}$$

où $GR1$ représente le groupe Setosa, $GR2$ représente le groupe Versicolor et $GR3$ le groupe Virginica. Ce jeu de données multivarié nous fournit donc un parfait exemple de classification auquel nous appliquons différentes méthodes. Afin de disposer également d'un jeu de données univarié, nous choisissons de prendre une des quatre composantes des observations précédentes. Le graphique 4.1.1

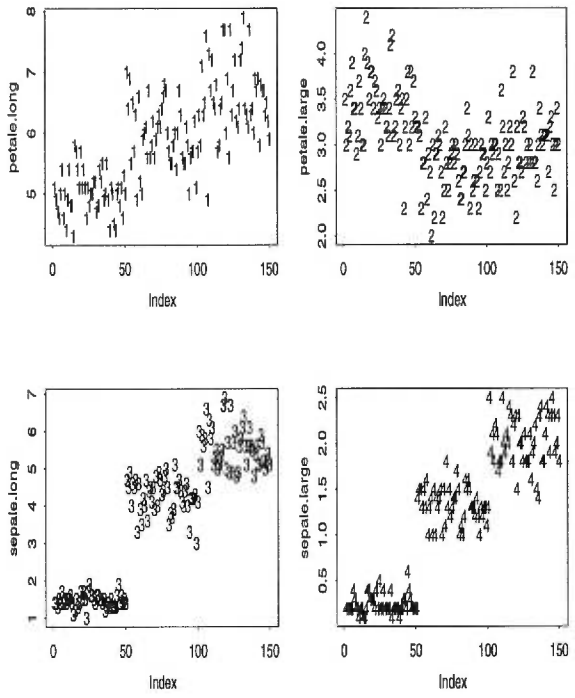

FIGURE 4.1.1. *Graphiques univariés des quatre composantes du jeu de données Iris*

nous permet de voir quelles sont les composantes dont la classification est la plus facilement identifiable. Ceci nous amène donc à choisir entre le troisième et le quatrième paramètre. Pour les fins de l'exemple, nous retenons le quatrième paramètre, c'est-à-dire la mesure de la largeur des pétales. Nous disposons donc de deux jeux de données, l'un multivarié et l'autre univarié.

4.2. Comparaison des différentes méthodes de classification

Les méthodes fréquentistes que nous nous proposons d'appliquer sont les suivantes :

- méthode du plus proche voisin,
- méthode du voisin le plus éloigné,
- méthode de la moyenne,
- méthode centroïde,
- méthode de la médiane,
- classification de Ward,
- méthode du maximum de vraisemblance (en supposant les données normales).

Toutes ces méthodes ont été présentées en détail dans le premier chapitre. Elles sont toutes disponibles sur différents logiciels statistiques, mais SAS a l'avantage de toutes les offrir en option. Afin d'avoir des résultats le plus facilement comparables, nous avons opté pour ce dernier logiciel et nous avons utilisé la procédure CLUST (voir SAS-User's guide) qui nous a permis d'observer les résultats que nous verrons plus tard. Toutes les méthodes citées ci-dessus étant des méthodes agglomératives, elles aboutissent toutes à la formation d'un seul groupe contenant toutes les observations. Dans le but de comparer ces résultats et ceux des méthodes de type bayésien, avec la classification " réelle ", nous choisissons de stopper les algorithmes lorsque nous obtenons trois groupes. Les méthodes d'optimisation fréquentistes sont considérées dans notre cas d'un point de vue bayésien (voir section 2.4). Parmi les trois approches découlant de ces méthodes, nous choisissons de minimiser la trace de la matrice W. Cette méthode n'étant

disponible sur aucun logiciel, nous avons donc dû la programmer. Pour des raisons de temps de calcul et de complexité évidentes, nous n'avons pas, bien sûr, considéré toutes les partitions possibles de 150 observations en 3 groupes (nous rappelons que le nombre de groupes est fixé à l'avance dans ce type d'approche). L'algorithme que nous avons choisi est le suivant (le programme se trouve en annexe B) :

i) classification initiale : trois données sont classées dans 3 groupes différents (la méthode centroïde, par exemple, peut être utilisée pour fournir trois observations appartenant à trois groupes différents). Le reste des données n'est pas classé.

ii) Pour chacune des n observations (incluant les trois observations initialement classées), nous calculons la trace de la matrice W lorsque l'observation est classée successivement dans les groupes $GR1$, $GR2$ et $GR3$. Nous choisissons alors la partition pour laquelle la trace est minimum.

Cet algorithme ne donne peut être pas la partition optimale, mais il a l'avantage d'être simple, efficace et rapide en termes de temps de calcul. Enfin, pour ce qui est de la méthode bayésienne présentée au troisième chapitre, seul l'application du test des moyennes (sans les variances), qui permet de regrouper les observations en classes à l'intérieur desquelles les moyennes sont identiques, nous permet une comparaison avec les autres méthodes. Les tests que nous effectuons sur les variances, quant à eux, raffinent les résultats et nous donnent des informations qui ne sont pas disponibles avec les autres méthodes. Nous verrons donc ces résultats plus tard. D'autre part, comme nous l'avons vu à la section 3.4, l'estimation du paramètre de nuisance se fait à l'aide de tableaux. Ceux-ci se présentent sous la forme (dans le cas univarié) du tableau 4.0.1. Nous avons calculé la probabilité *a posteriori* , p'_0 (voir section 3.3), en choisissant $X_1 = X_2 = 0$, pour plusieurs

TABLEAU 4.2.1. *Estimation des paramètre de nuisance : cas univarié*

α / n_0	0,00001	0,0001	0,001	0,01	0,1	1	1,5
1	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
1,5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
2	**0,9985491**	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
2,5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
3	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
3,5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
4	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
4,5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122
5,5	0,9985491	0,9954264	0,9856795	0,9561048	0,874006	0,7010649	0,6646122

valeurs de α et n_0. Nous ne choisissons pas alors forcement la combinaison (α, n_0) rendant maximum cette probabilité (les X_i étant égaux, il est normal d'avoir une probabilité *a posteriori* proche de 1), car nous voulons éviter de rendre la méthode trop " conservatrice ". En réalité, dans notre cas, quel que soit le choix de n_0, les résultats sont identiques au niveau du type et du nombre de groupes formés. Comme nous pouvons le constater dans le tableau, le paramètre α ne joue aucun rôle. Nous choisissons donc arbitrairement $\alpha = 2$ et nous posons $n_0 = 0{,}00001$. Le fait que n_0 soit très petit nous assure une très grande variance *a priori* des moyennes du modèle que nous avons choisi. Pour faire nos tests, nous comparons la cote de Bayes pour les moyennes (H_0' et H_1') à 1, pour les moyennes égales et variances différentes (H_0 et H_3) nous utilisons $c = 10$ afin de former de tels groupes. Enfin, pour comparer H_2 et H_3, nous utilisons $c = 1$.

Nous devrions présenter les résultats généraux de toutes les méthodes sous la forme de trois tableaux. Le premier contiendrait le classement des 50 premières données, le second celui des 50 suivantes et enfin, le dernier tableau contient le

classement des 50 dernières observations. En réalité, dans le cas des cinquante premières données, toutes les méthodes sont unanimes, que ce soit le cas univarié ou multivarié : le premier groupe est entièrement reconnu par toutes les méthodes. Cela signifie que les cinquante premières observations sont toutes classées dans le groupe 1. Un tableau décrivant ce type de résultat étant peu pertinent, nous choisissons donc de présenter uniquement deux tableaux, contenant la classification des données pour les deux derniers groupes. Dans chaque tableau, la première colonne contient le numéro d'observation, tandis que la deuxième nous donne le numéro du groupe auquel elle appartient réellement. Ensuite, chacune des colonnes suivantes indique le classement de l'observation selon une des méthodes étudiées. De plus, les chiffres en caractère gras représentent les erreurs de classement. Le code des méthodes est le suivant :

$$\begin{cases} RE : & \text{classement réel,} \\ CEN : & \text{méthode centroïde,} \\ MOY : & \text{méthode de la moyenne,} \\ VE : & \text{méthode du voisin le plus éloigné,} \\ ME : & \text{méthode de la médiane,} \\ VP : & \text{méthode du plus proche voisin,} \\ MV : & \text{méthode du maximum de vraisemblance,} \\ WD : & \text{méthode de Ward,} \\ OP : & \text{méthode d'optimisation,} \\ BAY : & \text{nouvelle approche bayésienne.} \end{cases}$$

TABLEAU 4.2.2: Classement du deuxième groupe : cas univarié

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
51	2	2	2	2	2	2	2	2	2	2
52	2	2	2	2	2	2	2	2	2	2
53	2	2	2	2	2	2	2	2	2	2
54	2	2	2	2	2	2	2	2	2	2
55	2	2	2	2	2	2	2	2	2	2
56	2	2	2	2	2	2	2	2	2	2
57	2	2	2	2	2	2	2	2	2	2
58	2	2	2	2	2	2	2	2	2	2
59	2	2	2	**3**	2	2	2	2	2	2
60	2	2	2	2	2	2	2	2	2	2
61	2	2	2	2	2	2	2	2	2	2
62	2	2	2	2	2	2	2	2	2	2
63	2	2	2	2	2	2	2	2	2	2
64	2	2	2	2	2	2	2	2	2	2
65	2	2	2	2	2	2	2	2	2	2
66	2	2	2	2	2	2	2	2	2	2
67	2	2	2	2	2	2	2	2	2	2
68	2	2	2	2	2	2	2	2	2	2
69	2	2	2	2	2	2	2	2	2	2
70	2	2	2	2	2	2	2	2	2	2
71	2	2	2	2	2	2	2	2	2	2
72	2	**3**	**3**	**3**	2	2	**3**	2	2	2
73	2	2	2	2	2	2	2	2	2	2
74	2	2	2	2	2	2	2	2	2	2
75	2	2	2	2	2	2	2	2	2	2
76	2	2	2	2	2	2	2	2	2	2
77	2	2	2	2	2	2	2	2	2	2
78	2	2	2	2	2	2	2	2	2	2
79	2	2	2	**3**	2	2	2	2	2	2
80	2	2	2	2	2	2	2	2	2	2

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
81	2	2	2	2	2	2	2	2	2	2
82	2	2	2	2	2	2	2	2	2	2
83	2	2	2	2	2	2	2	2	2	2
84	2	2	2	2	2	2	2	2	2	2
85	2	2	2	**3**	2	2	2	2	2	2
86	2	2	2	2	2	2	2	2	2	2
87	2	2	2	**3**	2	2	2	2	2	2
88	2	2	2	2	2	2	2	2	2	2
89	2	2	2	2	2	2	2	2	2	2
90	2	2	2	2	2	2	2	2	2	2
91	2	2	2	2	2	2	2	2	2	2
92	2	2	2	2	2	2	2	2	2	2
93	2	2	2	2	2	2	2	2	2	2
94	2	2	2	2	2	2	2	2	2	2
95	2	2	2	2	2	2	2	2	2	2
96	2	2	2	2	2	2	2	2	2	2
97	2	2	2	2	2	2	2	2	2	2
98	2	2	2	2	2	2	2	2	2	2
99	2	2	2	2	2	2	2	2	2	2
100	2	2	2	2	2	2	2	2	2	2

TABLEAU 4.2.3: Classement du troisième groupe : cas univarié

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
101	3	3	3	3	3	3	3	3	3	3
102	3	3	3	3	**2**	**2**	3	3	3	3
103	3	3	3	3	3	**2**	3	3	3	3
104	3	3	3	3	**2**	**2**	3	**2**	3	3
105	3	3	3	3	3	**2**	3	3	3	3
106	3	3	3	3	3	**2**	3	3	3	3

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
107	3	**2**	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**
108	3	3	3	3	**2**	**2**	3	**2**	3	3
109	3	3	3	3	**2**	**2**	3	**2**	3	3
110	3	3	3	3	3	3	3	3	3	3
111	3	3	3	3	3	**2**	3	3	3	3
112	3	3	3	3	**2**	**2**	3	3	3	3
113	3	3	3	3	3	**2**	3	3	3	3
114	3	3	3	3	3	**2**	3	3	3	3
115	3	3	3	3	3	3	3	3	3	3
116	3	3	3	3	3	**2**	3	3	3	3
117	3	3	3	3	**2**	**2**	3	**2**	3	3
118	3	3	3	3	3	**2**	3	3	3	3
119	3	3	3	3	3	**2**	3	3	3	3
120	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
121	3	3	3	3	3	**2**	3	3	3	3
122	3	3	3	3	3	**2**	3	3	3	3
123	3	3	3	3	3	**2**	3	3	3	3
124	3	3	3	3	**2**	**2**	3	**2**	3	3
125	3	3	3	3	3	**2**	3	3	3	3
126	3	3	3	3	**2**	**2**	3	**2**	3	3
127	3	3	3	3	**2**	**2**	3	**2**	3	3
128	3	3	3	3	**2**	**2**	3	**2**	3	3
129	3	3	3	3	3	**2**	3	3	3	3
130	3	**2**	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**
131	3	3	3	3	**2**	**2**	3	3	3	3
132	3	3	3	3	3	**2**	3	3	3	3
133	3	3	3	3	3	**2**	3	3	3	3
134	3	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**
135	3	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**
136	3	3	3	3	3	**2**	3	3	3	3

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
137	3	3	3	3	3	3	3	3	3	3
138	3	3	3	3	**2**	**2**	3	**2**	3	3
139	3	3	3	3	**2**	**2**	3	**2**	3	3
140	3	3	3	3	3	**2**	3	3	3	3
141	3	3	3	3	3	3	3	3	3	3
142	3	3	3	3	3	**2**	3	3	3	3
143	3	3	3	3	**2**	**2**	3	3	3	3
144	3	3	3	3	3	**2**	3	3	3	3
145	3	3	3	3	3	3	3	3	3	3
146	3	3	3	3	3	**2**	3	3	3	3
147	3	3	3	3	**2**	**2**	3	3	3	3
148	3	3	3	3	3	**2**	3	3	3	3
149	3	3	3	3	3	**2**	3	3	3	3
150	3	3	3	3	**2**	**2**	3	**2**	3	3

Au vu de ces résultats, nous pouvons établir un classement des méthodes, quand à leur performance. Ces résultats sont résumés dans le tableau 4.2.4. Tout d'abord, nous pouvons observer qu'une erreur commise par une méthode est souvent reprise par les autres. Cela signifie certainement que les observations mal classées contiennent quelque ambiguïté quant à leur classement réel. D'autre part, les méthodes se divisent en deux types : celles qui reconnaissent deux groupes distincts et celles qui ont tendance à homogénéiser les deuxièmes et troisièmes groupes. Les méthodes les plus performantes sont la nouvelle méthode bayésienne, les méthodes centroïde, de la moyenne, du voisin le plus éloigné, du maximum de vraisemblance et enfin, les méthodes d'optimisation. Dans ces cas-là, le pourcentage d'erreur varie de 3% à 4% (ce qui est équivalent à 4 à 8 observations mal classées). Le reste des méthodes ne distingue pas clairement les groupes 2 et 3. Le pourcentage d'erreur peut varier alors de 11% (méthode de Ward) à 30%

TABLEAU 4.2.4. *Pourcentage d'erreur avec le jeu de données univarié*

CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
4	3	5	14	29	4	11	3	3

(méthode du plus proche voisin). En réalité, c'est le niveau de " sévérité " des méthodes lors du regroupement de deux observations qui fait la différence dans cet exemple. Il semble donc que les méthodes qui performent le moins bien sont beaucoup trop conservatrices dans le cas d'un jeu de données comme celui que nous avons étudié. La nouvelle méthode obtient donc d'excellents résultats dans le cas univarié.

Pour le cas multivarié, nous présentons les résultats de la même façon, c'est-à-dire sous la forme de trois tableaux. En ce qui concerne les paramètres de nuisance de la nouvelle méthode, nous estimons α et n_0 de la même façon que pour le cas univarié. Les valeurs du paramètre α sont plus élevées que dans le cas précédent. Ceci est dû tout simplement au fait que les fonctions gamma multivariées que nous calculons dans notre algorithme ne sont pas définies pour les petites valeurs de α.

Le tableau 4.2.5 nous permet de choisir $\alpha = 4$ et $n_0 = 0{,}1$. Une valeur de α trop petite aurait rendu la méthode trop conservatrice. Nous avons donc choisi une valeur de n_0 moins " extrême ". Cette fois-ci, nous pouvons constater que, contrairement au cas univarié, le paramètre α semble avoir plus d'effet sur les valeurs de la densité marginale. Cependant, en autant que nous ne choisissions pas des valeurs de α trop petites, nous avons pu constater que quelle que soit la valeur de α ou de n_0, les résultats restent identiques (voir tableau 4.2.6).

TABLEAU 4.2.5. *Estimation des paramètres de nuisance : cas multivarié*

α / n_0	0,001	0,01	0,1	1	1,5	2
4	1	0.9999980	**0.9998059**	0.9851630	0.9711256	0.9553999
4,5	1	0.9999986	0.9998620	0.9894038	0.9792944	0.9678704
5	1	0.9999989	0.9998965	0.9920318	0.9843900	0.9757077
5,5	1	0.9999992	0.9999193	0.9937800	0.9877943	0.9809689
6	1	0.9999993	0.9999353	0.9950049	0.9901863	0.9846777
6,5	1	0.9999995	0.9999469	0.9958978	0.9919335	0.9873932
7	1	0.9999995	0.9999556	0.9965694	0.9932498	0.9894425
7,5	1	0.9999996	0.9999624	0.9970877	0.9942667	0.9910277
8	1	0.9999997	0.9999676	0.9974962	0.9950689	0.9922797
8,5	1	0.9999997	0.9999719	0.9978240	0.9957132	0.9932860

TABLEAU 4.2.6: Classement du deuxième groupe : cas multivarié

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
51	2	2	2	2	2	2	2	2	2	2
52	2	2	2	**3**	2	2	**3**	2	2	2
53	2	2	2	**3**	2	2	**3**	2	2	2
54	2	2	2	**3**	2	2	**3**	2	2	2
55	2	2	2	2	2	2	2	2	2	2
56	2	2	2	**3**	2	2	**3**	2	2	2
57	2	2	2	2	2	2	2	2	2	2
58	2	2	2	**3**	2	2	**3**	2	2	2
59	2	2	2	2	2	2	2	2	2	2
60	2	2	2	**3**	2	2	**3**	2	2	2
61	2	2	2	2	2	2	2	2	2	2
62	2	2	2	2	2	2	2	2	2	2
63	2	2	2	2	2	2	**3**	2	2	2
64	2	2	2	2	2	2	2	2	2	2

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
65	2	2	2	**3**	2	2	**3**	2	2	2
66	2	2	2	2	2	2	2	2	2	2
67	2	2	2	**3**	2	2	**3**	2	2	2
68	2	2	2	2	2	2	2	2	2	2
69	2	2	2	2	2	2	2	2	2	2
70	2	2	2	**3**	2	2	2	2	2	2
71	2	2	2	2	2	2	2	2	2	2
72	2	2	2	**3**	2	2	**3**	2	2	2
73	2	2	2	2	2	2	**3**	2	2	2
74	2	2	2	**3**	2	2	**3**	2	2	2
75	2	2	2	**3**	2	2	**3**	2	2	2
76	2	2	2	**3**	2	2	**3**	2	2	2
77	2	2	2	**3**	2	2	**3**	2	2	2
78	2	2	2	**3**	2	2	**3**	2	**3**	2
79	2	2	2	**3**	2	2	**3**	**3**	2	2
80	2	2	2	**3**	2	2	**3**	2	2	2
81	2	2	2	2	2	2	2	2	2	2
82	2	2	2	2	2	2	2	2	2	2
83	2	2	2	2	2	2	2	2	2	2
84	2	2	2	2	2	2	2	2	**3**	2
85	2	2	2	**3**	2	2	**3**	2	2	2
86	2	2	2	2	2	2	2	2	2	2
87	2	2	2	**3**	2	2	**3**	2	**3**	2
88	2	2	2	**3**	2	2	**3**	2	2	2
89	2	2	2	**3**	2	2	2	2	2	2
90	2	2	2	2	2	2	2	2	2	2
91	2	2	2	2	2	2	2	2	2	2
92	2	2	2	2	2	2	2	2	2	2
93	2	2	2	**3**	2	2	**3**	2	2	2
94	2	2	2	2	2	2	2	2	2	2

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
95	2	2	2	2	2	2	2	2	2	2
96	2	2	2	2	2	2	2	2	2	2
97	2	2	2	2	2	2	2	2	2	2
98	2	2	2	2	2	2	2	2	2	2
99	2	2	2	**3**	2	2	**3**	2	2	2
100	2	2	2	2	2	2	2	2	2	2

TABLEAU 4.2.7: Classement du troisième groupe : cas multivarié

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
101	3	3	**2**	3	3	**2**	3	3	3	3
102	3	**2**	**2**	3	**2**	**2**	3	**2**	3	**2**
103	3	3	3	3	3	**2**	3	3	3	3
104	3	3	**2**	3	**2**	**2**	3	3	3	3
105	3	3	**2**	3	**2**	**2**	3	3	3	3
106	3	3	3	3	3	**2**	3	3	3	3
107	3	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
108	3	3	3	3	3	**2**	3	3	3	3
109	3	3	**2**	3	**2**	**2**	3	3	3	3
110	3	3	3	3	3	**2**	3	3	3	3
111	3	3	**2**	3	**2**	**2**	3	3	3	3
112	3	3	**2**	3	**2**	**2**	3	3	3	3
113	3	3	**2**	3	**2**	**2**	3	3	3	3
114	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
115	3	**2**	**2**	3	**2**	**2**	3	**2**	3	**2**
116	3	3	**2**	3	**2**	**2**	3	3	3	3
117	3	3	**2**	3	**2**	**2**	3	3	3	3
118	3	3	3	3	3	3	3	3	3	3
119	3	3	3	3	3	**2**	3	3	3	3
120	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**

OBS	RE	CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
121	3	3	**2**	3	**2**	**2**	3	3	3	3
122	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
123	3	3	3	3	3	**2**	3	3	3	3
124	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
125	3	3	**2**	3	**2**	**2**	3	3	3	3
126	3	3	3	3	3	**2**	3	3	3	3
127	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
128	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
129	3	3	**2**	3	**2**	**2**	3	3	3	3
130	3	3	3	3	3	**2**	3	3	3	3
131	3	3	3	3	3	**2**	3	3	3	3
132	3	3	3	3	3	3	3	3	3	3
133	3	3	**2**	3	**2**	**2**	3	3	3	3
134	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
135	3	3	**2**	3	**2**	**2**	3	**2**	3	3
136	3	3	3	3	3	**2**	3	3	3	3
137	3	3	**2**	3	**2**	**2**	3	3	3	3
138	3	3	**2**	3	**2**	**2**	3	3	3	3
139	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
140	3	3	**2**	3	**2**	**2**	3	3	3	3
141	3	3	**2**	3	**2**	**2**	3	3	3	3
142	3	3	**2**	3	**2**	**2**	3	3	3	3
143	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
144	3	3	**2**	3	**2**	**2**	3	3	3	3
145	3	3	**2**	3	**2**	**2**	3	3	3	3
146	3	3	**2**	3	**2**	**2**	3	3	3	3
147	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**
148	3	3	**2**	3	**2**	**2**	3	3	3	3
149	3	3	**2**	3	**2**	**2**	3	3	3	3
150	3	**2**	**2**	3	**2**	**2**	3	**2**	**2**	**2**

Cette fois-ci, dans le cas multivarié, les résultats sont un peu différents. Nous pouvons analyser le pourcentage d'erreur à l'aide du tableau 4.2.8. Les quatre méthodes qui performent le mieux sont la nouvelle méthode bayésienne et la méthode d'optimisation (qui était déjà performante dans le cas univarié), la méthode centroïde et la méthode de Ward. Les pourcentages d'erreur se situent alors entre 9 % et 12 %. Contrairement au cas précédent, nous pouvons constater que la méthode de Ward est meilleure dans le cas multivarié. Les méthodes ayant le plus de difficultés à reconnaître les trois groupes sont la méthode de la moyenne, la méthode du maximum de vraisemblance (qui obtenaient toutes deux pourtant de très bons résultats dans le cas univarié) et la méthode de la médiane (qui dans les deux cas ne performe pas très bien). Les taux d'erreur dans ces cas-là varient de 25 % à 32 %. Comme précédemment, il semble que certaines approches distinguent mal la frontière entre les deuxième et troisième groupes. Ceci peut plus facilement s'expliquer dans le cas multivarié car nous pouvons voir sur le graphique (4.1.1) que la première composante des variables ne se divise pas de façon claire en trois groupes et que la deuxième composante, par contre, se divise nettement en deux classes et non trois. Bien sûr, ceci expliquerait le taux d'erreur, plus élevé avec le jeu de données multivarié. Nous pouvons donc conclure que l'influence de ces deux composantes est assez forte dans le cas des méthodes qui ne performent pas comme nous l'aurions souhaité. Pour ce qui est de la méthode du plus proche voisin, qui affiche 48 observations mal classées, nous rappelons que cette méthode calcule la distance entre deux groupes à partir des deux observations les plus proches. La frontière entre les groupes 2 et 3 n'étant pas clairement définie, il est normal que dans ce cas, cette méthode ne donne pas de résultats satisfaisants.

Pour résumer, trois méthodes semblent performer particulièrement bien dans le cas des iris de Fisher : la méthode centroïde, la méthode d'optimisation et

TABLEAU 4.2.8. *Pourcentage d'erreur avec le jeu de données multidimensionnel*

CEN	MOY	VE	ME	VP	MV	WD	OP	BAY
9	25	13	24	32	17	11	10	9

la nouvelle méthode bayésienne. Cette dernière est d'ailleurs la seule à obtenir les meilleurs résultats dans les deux cas, multivarié et univarié. Les méthodes bayésiennes obtiennent donc de très bons résultats, particulièrement dans le cas multivarié. Nous pouvons noter que parmi toutes les méthodes que nous avons présentées, celle que nous avons développée est la seule dont le nombre de groupes n'est pas fixé à l'avance. Bien sûr, le fait d'estimer les paramètres de nuisance à l'aide de tableaux permet de jouer sur le degré de " conservatisme " de la méthode, mais en réalité, dans le cas univarié aussi bien que dans le cas multivarié, quel que soit le coefficient choisi, nous aboutissons toujours à trois groupes. Les résultats sont de plus excellents, et nous pouvons noter que cette dernière approche est la seule à reconnaître les trois types d'iris de façon " naturelle ", ce qui lui confère un net avantage.

4.3. CLASSIFICATION PAR LA VARIANCE

Outre la classification par la moyenne, la méthode bayésienne que nous avons développé fournit aussi une classification par la variance, comme nous l'avons vu au troisième chapitre. Dans le cas des iris de Fisher, le test d'égalité ou non des variances, sachant les moyennes égales, ne s'avère pas très concluant. En effet, dans les tests de moyennes, décrits au troisième chapitre, nous supposons implicitement les variances des observations égales entre elles. Ceci nous amène donc à la formation de groupes dont les observations les composant sont assez concentrées autour de la moyenne de la classe. Par ce fait même, il est impossible de détecter

une différence substantielle au niveau des variances et donc, aucun sous-groupe ne se forme à l'intérieur des groupes. Par contre, les résultats concernant l'égalité de variances entre les trois groupes formés sont beaucoup plus intéressants. Pour ce qui est du cas univarié, nous aboutissons à une égalité de variances des deuxième et troisième groupes. Graphiquement, ce résultat était prévisible. En effet, si l'on observe encore une fois le graphique 4.1.1 et en particulier la quatrième composante que nous avons étudiée, nous pouvons constater que la forme des deux derniers groupes est assez semblable, contrairement au premier groupe. Il n'est donc pas tellement surprenant que notre algorithme aboutisse à une égalité de variances des groupes 2 et 3. Dans le cas multivarié, les résultats nous amènent à considérer les trois groupes formés comme ayant la même variance.

Les résultats que nous obtenons sur le jeu de données des iris de Fisher ne sont pas forcément représentatifs de la " qualité " des méthodes en général. En effet, nous devons comprendre que certaines méthodes sont meilleures que d'autres dans le cas d'un jeu de données ayant une forme particulière. D'autres performent mieux dans d'autres situations. Certaines méthodes, comme celle du maximum de vraisemblance par exemple, sont même développées dans un cadre bien précis (les modèles mixtes, en l'occurrence) et ne donnent pas forcément de très bons résultats lorsque les observations ne suivent pas de loi normale. Le but de ce chapitre n'était donc pas de conclure quant à la qualité des méthodes d'un point de vue général.

CONCLUSION

Le but de ce mémoire est donc de développer une nouvelle méthode bayésienne de classification hiérarchique agglomérative, et de la comparer aux méthodes usuelles fréquentistes et bayésiennes. Les méthodes fréquentistes que nous avons choisi d'étudier sont principalement celles que l'on retrouve couramment dans la plupart des logiciels statistiques, notamment la méthode du plus proche voisin, celle du voisin le plus éloigné, la méthode de la moyenne, de la médiane, la méthode centroïde, la classification de Ward, les méthodes d'optimisation et celle du maximum de vraisemblance. Toutes ces méthodes ont des caractéristiques différentes mais dans la plupart des cas, elles diffèrent par le type de distance choisie. La méthode que nous avons choisie de présenter et de développer s'appuie sur des arguments bayésiens, puisque nous considérons les paramètres de notre modèle comme étant aléatoires et donc, nous utilisons l'information *a priori* que nous avons sur les données. Alors que beaucoup de méthodes usuelles de classification utilisent différentes distances comme critère de regroupement, nous avons choisi de construire différents tests d'hypothèses, utilisant la cote de Bayes, et nous permettant de tester l'homogénéité de deux classes d'observations que nous cherchons éventuellement à regrouper ensemble. Dans notre cas, l'homogénéité que nous cherchons à obtenir à l'intérieur de chaque classe n'implique pas seulement les valeurs des observations, mais aussi leur variance. En effet, nous cherchons à raffiner le type de classification usuel en introduisant une dimension de variabilité

à l'intérieur des groupes. Ces derniers peuvent donc être de différents type, selon que la moyenne ou la variance à l'intérieur des classes est identique ou différente.

La comparaison avec les autres méthodes de classification s'effectue dans notre cas sur des groupes à l'intérieur desquels nous avons accepté l'hypothèse d'égalité des moyennes. Les résultats que nous obtenons sont excellents. Nous avons appliqué toutes ces méthodes sur le jeu de données des iris de Fisher, que nous avons utilisé de façon univariée et multivariée. Ce jeu de données a l'avantage de nous fournir une classification pré-établie des données en trois groupes. C'est donc à cette classification que nous comparons toutes les méthodes. Dans le cas univarié comme multivarié, nous avons pu constater que la nouvelle méthode bayésienne est celle qui performe le mieux parmi celles utilisées couramment en classification. Le taux d'erreur est cependant plus élevé dans le cas multidimensionnel. Ceci est dû au fait que la frontière " réelle " entre le deuxième et troisième groupe n'est pas clairement établie. De plus, la nouvelle approche a le très gros avantage de ne pas fixer le nombre de groupes à l'avance, ni de poser une information *a priori* sur celui-ci. Les trois types d'iris sont donc reconnus de façon entièrement naturelle.

Il est malheureusement difficile en classification de généraliser sur la performance des méthodes utilisées, car cela dépend trop du type de jeu de données choisi. Certaines méthodes, qui performent très mal dans notre situation, s'avèrent excellentes dans d'autres cas. La méthode que nous avons développée possède de très gros avantages techniques par rapport aux autres approches. Nous pourrions toutefois l'améliorer au niveau de la complexité de calcul et du temps de compilation des programmes. En conclusion, nous pouvons donc affirmer que la nouvelle méthode bayésienne peut se comparer avantageusement aux méthodes usuelles classiques, même si elle s'avère beaucoup plus complexe en termes de calculs.

Annexe A

ANNEXE A : LES PROGRAMMES MULTIVARIÉS

```
##############################################################################
# PROGRAMME DE REGROUPEMENT PAR MOYENNE EGALE, PUIS PAR VARIANCES EGALES #
# APPROCHE MULTIDIMENSIONNELLE                                               #
##############################################################################



# ESTIMATION DES PARAMETRES ALPHA, BETA, N0 ET TETA0
# --------------------------------------------------

estimpara_function(n1,n2,S1,S2,x1,x2)
{
   teta0_(n1*x1+n2*x2)/(n1+n2)
   alpha_4
   B_diag(p)
   n0_0.1
   list("n0"=n0,"alpha"=alpha,"B"=B,"teta0"=teta0)
}

# DETERMINANT
det_function(A)
{
        c <- eigen(A)$values
        det <- prod(c)
        return(det)
}

# FONCTION GAMMA MULTIVARIE
gamp_function(a,p)
{
```

```
   produit_1
   for (i in 1:p)
   {
       produit_produit*exp(lgamma(a-0.5*(i-1)))
   }
   gamp_pi^(p*(p-1)/4)*produit
   return(gamp)
}

# TEST MOYENNES EGALES : M0'
m0mult.moy_function(p,n0,n1,n2,B,alpha,teta0,x1,x2)
{
  cste0_(n0*n1*n2)^(p/2)/((2*pi)^p)*det(B)^(alpha/2)/gamp(alpha/2,p)/
        2^(alpha*p/2)
  cste1_gamp(1+alpha/2,p)/(n0+n1+n2)^(p/2)
  den_B+n1*n2/(n0+n1+n2)*(x1-x2)%*%t(x1-x2)+n0*n1/(n0+n1+n2)*
     (x1-teta0)%*%t(x1-teta0)+ n0*n2/(n0+n1+n2)*(x2-teta0)%*%t(x2-teta0)
  multi.m0_cste0*cste1/(det(den))^(alpha/2+1)
  return(multi.m0)
}

# TEST MOYENNES DIFFERENTES : M1'
m1mult.moy_function(p,n0,n1,n2,B,alpha,teta0,x1,x2)
{
   cste0_(n0*n1*n2)^(p/2)/((2*pi)^p)*det(B)^(alpha/2)/gamp(alpha/2,p)/
         2^(alpha*p/2)
   cste1_n0^(p/2)/((n0+n1)*(n0+n2))^(p/2)*gamp((alpha+1)/2,p)
   den_B+n0*n1/(n0+n1)*(x1-teta0)%*%t(x1-teta0)+n0*n2/(n0+n2)*
       (x2-teta0)%*%t(x2-teta0)
   multi.m1_cste0*cste1/(det(den))^((alpha+1)/2)
   return(multi.m1)
}

# TEST MOYENNES EGALES, VARIANCES EGALES : M0
m0mult.moyeg.vareg_function(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
{
#   cste0_(n0*n1*n2)^(p/2)/(2*pi)^p*det(B)^(alpha/2)/gamp(alpha/2,p)*
#         det(S1)^((n1-p)/2-1)*det(S2)^((n2-p)/2-1)/gamp((n1-1)/2,p)/
#         gamp((n2-1)/2,p)/2^(alpha*p/2)/2^(p*((n1+n2)/2-1))

   cste1_gamp((n1+n2+alpha)/2,p)/(n0+n1+n2)^(p/2)
#   cste1_aide.calcul.m0() /(n0+n1+n2)^(p/2)

   den_B+S1+S2+n1*n2/(n0+n1+n2)*(x1-x2)%*%t(x1-x2)+n0*n1/(n0+n1+n2)*
       (x1-teta0)%*%t(x1-teta0)+n0*n2/(n0+n1+n2)*(x2-teta0)%*%t(x2-teta0)
   multi.m0_cste1/(det(den))^((n1+n2+alpha)/2)
```

```
   return(multi.m0)
}

# TEST MOYENNES DIFFERENTES, VARIANCES DIFFERENTES : M1
m1mult.moydiff.vardiff_function(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
{
#   cste0_(n0*n1*n2)^(p/2)/(2*pi)^p*det(B)^(alpha/2)/gamp(alpha/2,p)*
#         det(S1)^((n1-p)/2-1)*det(S2)^((n2-p)/2-1)/gamp((n1-1)/2,p)/
#         gamp((n2-1)/2,p)/2^(alpha*p/2)/ 2^(p*((n1+n2)/2-1))

   cste1_n0^(p/2)/((n0+n1)*(n0+n2))^(p/2)*det(B)^(alpha/2)/gamp(alpha/2,p)/
         2^(alpha*p/2)*gamp((alpha+n1-1)/2,p)*gamp((alpha+n2-1)/2,p)*
         2^(p*(2*alpha+n1+n2-2))

   den1_B+S1+n0*n1/(n0+n1)*(x1-teta0)%*%t(x1-teta0)
   den2_B+S2+n0*n2/(n0+n2)*(x2-teta0)%*%t(x2-teta0)
   multi.m1_cste1/(det(den1))^((n1+alpha-1)/2)/(det(den2))^((n2+alpha-1)/2)
   return(multi.m1)
}

# TEST MOYENNES DIFFERENTES, VARIANCES EGALES : M2
m2mult.moydiff.vareg_function(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
{
#   cste0_(n0*n1*n2)^(p/2)/(2*pi)^p*det(B)^(alpha/2)/gamp(alpha/2,p)*
#         det(S1)^((n1-p)/2-1)*det(S2)^((n2-p)/2-1)/gamp((n1-1)/2,p)/
#         gamp((n2-1)/2,p)/2^(alpha*p/2)/2^(p*((n1+n2)/2-1))

   cste1_gamp((n1+n2+alpha)/2,p)/((n0+n1)*(n0+n2))^(p/2)*2^(p*(n1+n2+alpha))

   den_B+S1+S2+n0*n1/(n0+n1)*(x1-teta0)%*%t(x1-teta0)+
       n0*n2/(n0+n2)*(x2-teta0)%*%t(x2-teta0)
   multi.m2_cste1/(det(den))^((n1+n2+alpha)/2)
   return(multi.m2)
}

# CALCUL DE LA TRACE D'UNE MATRICE
trace.mat_function(A,n)
{
   trace.mat_0
   for (i in 1:n)
   {
      trace.mat_trace.mat+A[i,i]
   }
   return(trace.mat)
}
```

```
# TEST MOYENNES EGALES, VARIANCES DIFFERENTES : M3
 m3mult.moyeg.vardiff_function(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
{
   iter_500
#   cste0_(n0*n1*n2)^(p/2)/(2*pi)^p*det(B)^(alpha/2)/gamp(alpha/2,p)*
#          det(S1)^((n1-p)/2-1)*det(S2)^((n2-p)/2-1)/gamp((n1-1)/2,p)/
#          gamp((n2-1)/2,p)/2^(alpha*p/2)/2^(p*((n1+n2)/2-1))


   gam_0
   if ( (2*alpha+p+n1+n2-1)/2>58)  {gam_gamp(58,p)}
   else {gam_gamp((2*alpha+p+n1+n2-1)/2,p)}
   cste1_1/(1+n0)^(p/2)*det(B)^(alpha/2)/gamp(alpha/2,p)*
          gam/200^(p*(2*alpha+p+n1+n2-1)/2)*2^(p*(alpha/2+p-1))*
          gamma(alpha+n2)
   den_vector(mode="numeric",length=iter)
   num_vector(mode="numeric",length=iter)
   result_vector(mode="numeric",length=iter)
   A_matrix(0,p,p)
   for (i in 1:iter)
   {
      v_rgamma(1,(alpha+n2)/2)
      v_1/v
      A_1/(n1*v+n2)*(n1*v*x1+n2*x2)
      I_diag(p)
      den1_B*(v+1)/v+S1+S2/v+n1*n2/(n1*v+n2)*(x1-x2)%*%t(x1-x2)+
           1/v*(n1*v+n2)*(teta0-A)%*%t(teta0-A)*n0/(n0+1)

      den[i]_det(den1/200)^((2*alpha+p+n1+n2-1)/2)
      num[i]_exp(1/v)*v^((alpha+n2+1-p)/2)
      result[i]_num[i]/den[i]
   }
   m3.multi_cste1*mean(result)
   return(m3.multi)
}

# CALCUL DE P0' ET DE P1'
cotemult.moy_function(p,x1,x2, n1, n2)
{
   teta0_(n1*x1+n2*x2)/(n1+n2)
   alpha_4
   B_diag(p)
   M0_m0mult.moy(p,n0,n1,n2,B,alpha,teta0,x1,x2)
   M1_m1mult.moy(p,n0,n1,n2,B,alpha,teta0,x1,x2)
   S_M0+M1
   p0_M0/S
```

```
   p1_M1/S
   list("p0"=p0,"p1"=p1,"cote"=M0/M1)
}

# CALCUL DE P0 ET P3
cotemult.moyeg_function(p,x1,x2, n1, n2, S1,S2)
{
   result_estimpara(n1,n2,S1,S2,x1,x2)
   n0_result$n0
   alpha_result$alpha
   B_result$B
   teta0_result$teta0
   M0_m0mult.moyeg.vareg(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
   M3_m3mult.moyeg.vardiff(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
   S_M3+M0
   p0_M0/S
   p3_M3/S
 #  list("p0"=p0,"p1"=p3,"cote"=M0/M3)
   return(M0/M3)
}

# CALCUL DE P2 ET P1
cotemult.moydif_function(p,x1,x2, n1, n2, S1,S2)
{
   result_estimpara(n1,n2,S1,S2,x1,x2)
   n0_result$n0
   alpha_result$alpha
   B_result$B
   teta0_result$teta0
   M2_m2mult.moydiff.vareg(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
   M1_m1mult.moydiff.vardiff(p,n0,n1,n2,B,alpha,teta0,x1,x2,S1,S2)
   S_M1+M2
   p0_M2/S
   p1_M1/S
   list("p0"=p0,"p1"=p1,"cote"=M2/M1)
}

# CALCUL DE LA DISTANCE ENTRE 2 POINTS
d_function(x,y)
{
   d_sum((x-y)^2)
   return(d)
}

# TESTE SI UN VECTEUR EST VIDE ENTIEREMENT
vide.vect_function(x)
```

```
{
   result_T
   n_length(x)
   for (i in 1:n)
   {
      if(is.na(x[i])==F){result_F}
   }
   return(result)
}

# CREATION DE LA MATRICE TRIANGULAIRE SUPERIEURE DE DISTANCES
Dist_function(X,n)
{
   M_matrix(0,n,n)
   for (i in 1:(n-1))
   {
      for (j in (i+1):n)
      {
         if (vide.vect(X[,i])==F & vide.vect(X[,j])==F)
         {
            M[i,j]_d(X[,i],X[,j])
         }
         if (vide.vect(X[,i])==T | vide.vect(X[,j])==T)
         {
            M[i,j]_ NA
         }
      }
   }
    return(M)
 }

# TROUVE LE MIN DE LA MATRICE DE DISTANCE ET DETERMINE SA POSITION
trouvmin_function(M,n)
{
   mini_max(M,na.rm=T)+1
   i0_1
   j0_2
   for (i in 1:(n-1))
   {
      for (j in (i+1):n)
      {
         if(M[i,j]<mini & is.na(M[i,j])==F)
         {
            mini_M[i,j]
            i0_i
            j0_j
```

```
        }
      }
   }
   list("min"=mini,"i0"=i0,"j0"=j0)
}

# RETOURNE QUELLES SONT LES OBSERVATIONS QUI FORMENT LE GROUPE NUMERO K
formgroup_function(k,groupe,n)
{
   j_1
   vectgroup_rep.int(0,n)
   for (i in 1:n)
   {
      if (groupe[i]==k)
      {
         vectgroup[j]_i
         j_j+1
      }
   }
   return (vectgroup)
}

# FONCTION IS.NA APPLIQUEE A UNE MATRICE
mat.vide_function(M,n)
{
   resul_T
   test_1
   for (i in 1:n)
   {
      for (j in 1:n)
      {
         if ((is.na(M[i,j])==T) | (M[i,j]==0)) {test_test*1}
         else {test_test*0}
      }
   }
   if (test==0){resul_F}
   else{resul_T}
   return(resul)
}

# CALCUL DE SI ET DE XBARRE POUR UN VECTEUR VECTGROUP
cal.var_function(p,vectgroup,donnees)
{
  i_1
  xb_rep.int(0,p)
  while (vectgroup[i]!=0)
```

```
  {
     xb_xb+donnees[,vectgroup[i]]
     i_i+1
  }
  xb_xb/taille(vectgroup)
  xb_t(t(xb))
  i_1
  S_matrix(0,p,p)
  while (vectgroup[i]!=0)
  {
     val_donnees[,vectgroup[i]]
     val_t(t(val))
     S_S+(val-xb)%*%t(val-xb)
     i_i+1
  }
  return(S)
}

# TAILLE DU GROUPE VECTGROUP
taille_function(vectgroup)
{
   nbre_0
   i_1
   while (vectgroup[i]!=0)
   {
      nbre_nbre+1
      i_i+1
   }
   return(nbre)
}

# TROUVE LE MAX DANS UN TABLEAU S ENTRE LA DIFFERENCE S0-S[I] POUR TOUT I
#    OU S0 ET S[I] SONT EN FAIT DES TRACES DE MATRICES
trouvmax_function(s0,s)
{
   dmax_d(s0,s[1])
   i_2
   i0_1
   while (s[i]!=-1)
   {
      if (d(s0,s[i])>dmax)
      {
         dmax_d(s0,s[i])
         i0_i
      }
      i_i+1
```

```
   }
   return(i0)
}

# CALCULE LA MOYENNE D'UN GROUPE VECTGROUP EN OTANT OU NON LA K-IEME OBS
calmoy_function(p,k,vectgroup)
{
   if(k==0)
   {
      s_rep.int(0,p)
      i_1
      while(vectgroup[i]!=0)
      {
         s_s+donnees[,vectgroup[i]]
         i_i+1
      }
      result_s/(taille(vectgroup))
   }
   else
   {
      s_rep.int(0,p)
      i_1
      while(vectgroup[i]!=0)
      {
         if(i!=k)
         {
            s_s+donnees[vectgroup[i]]
         }
         i_i+1
      }
      result_s/(taille(vectgroup)-1)
   }
   return(result)
}

# TROUVE LES DEUX OBS LES PLUS PROCHES ET TESTE SI ON PEUT LES REGROUPER
# CETTE FONCTION RETOURNE LE VECTEUR GROUPE MODIFIE OU NON
corps_function(p,n,X,groupe,poids,vectgroup,nbgroup,M, n1, n2, donnees,
               arret,bic,cote.moyeg)
{
   vectgroup_rep.int(0,n)
   m_trouvmin(M,n)
   i0_m$i0
   j0_m$j0
   Y1_t(t(X[,i0]))
   Y2_t(t(X[,j0]))
```

```
   n1_poids[i0]
   n2_poids[j0]
   cote_cotemult.moy(p,Y1,Y2, n1, n2)$cote
   if (cote>cote.moyeg)
   {
         a_min(groupe[i0],groupe[j0])
         b_groupe[j0]
         e_groupe[i0]
         for (i in 1:n)
         {
            if (groupe[i]==e | groupe[i]==b)
            {
               groupe[i]_a
            }
         }
         vectgroup_formgroup(a,groupe,n)
         X[,i0]_(n1*Y1+n2*Y2)/(n1+n2)
         X[,j0]_t(t(rep.int(NA,p)))
         poids[i0]_poids[i0]+poids[j0]
         poids[j0]_0
         M[j0,]_NA
         M[i0,]_NA
         M[,j0]_NA
         M[,i0]_NA
         nbgroup_nbgroup-1
   }
   list("vectgroup"=vectgroup,"groupe"=groupe,"X"=X,"poids"=poids,
        "M"=M,"nbgroup"=nbgroup,"arret"=arret)

}

# CREATION DES GROUPES PAR MOYENNE EGALE
pro_function(p,n,X,groupe,poids,vectgroup,M, n1, n2,donnees,arret,cote.moyeg)
{
   arret_F
   group.plus_T
   nbgroup_n
   vectgroup_rep.int(0,n)
   vectgroup[1]_1
   while ((group.plus==T) & (mat.vide(M,n)==F) & (arret==F))
   {
      result_corps (p,n,X,groupe,poids,vectgroup,nbgroup,M, n1, n2,donnees,
                    arret,bic,cote.moyeg)
      vectgroup_result$vectgroup
      groupe_result$groupe
      X_result$X
```

```
      poids_result$poids
      M_result$M
      nbgroup_result$nbgroup
      arret_result$arret
      if (vectgroup[1]!=0) {group.plus_T}
      else {group.plus_F}
      while ((vectgroup[1]!=0) & (mat.vide(M,n)==F) & (arret==F))
      {
         result_corps (p,n,X,groupe,poids,vectgroup,nbgroup,M, n1, n2,donnees,
                       arret,bic,cote.moyeg)
         vectgroup_result$vectgroup
         groupe_result$groupe
         X_result$X
         poids_result$poids
         M_result$M
         nbgroup_result$nbgroup
         arret_result$arret
      M_Dist(X,n)
      }
   }
   list("vectgroup"=vectgroup,"groupe"=groupe,"X"=X,"poids"=poids,"M"=M,
        "nbgroup"=nbgroup,"arret"=arret)
}

# TRANSFORME GROUPE POUR AVOIR DES NUM DE 1 A NBGROUPE
transgr_function(groupe,n)
{
   tabmin_sort(groupe)
   fin_n
   i_1
   while (i<=(fin-1))
   {
      if(tabmin[i]==tabmin[i+1])
      {
         tabmin_tabmin[-(i+1)]
         fin_fin-1
         i_i-1
      }
      i_i+1
   }
   for (i in 1:fin)
   {
      for (j in 1:n)
      {
         if(groupe[j]==tabmin[i])
         {
```

```
            groupe[j]_i
         }
      }
   }
   return(groupe)
}

# CONSTRUCTION D'UN SOUS GROUPE DE VARIANCE EGALE A L'INTERIEUR D'UN GROUPE
corps2_function(p,g,groupe,n,donnees,nbssgroup,vectgroup,ssgroup,
                cote.moyeg.vardif)
{
   donnee.plus_T
   ssgroup.plus_F
   vectgroup.reste_rep.int(0,n)
   ctr_1
   while((donnee.plus==T) & (taille(vectgroup)>2) )
   {
      S_rep.int(-1,n)
      S0_trace.mat(cal.var(p,vectgroup,donnees),p)
      i_1
      while (vectgroup[i]!=0)
      {
         S[i]_trace.mat(cal.var(p,vectgroup[-i], donnees),p)
         i_i+1
      }
      i0_trouvmax(S0,S)
      y1_calmoy(p,0,vectgroup)
      y1_t(t(y1))
      y2_calmoy(p,i0,vectgroup)
      y2_t(t(y2))
      n1_taille(vectgroup)
      n2_taille(vectgroup)-1
      S1_cal.var(p,vectgroup,donnees)
      S2_cal.var(p,vectgroup[-i0],donnees)
      cote_cotemult.moyeg(p,y1,y2, n1, n2, S1,S2)
      if (cote<cote.moyeg.vardif)
      {
         ssgroup[vectgroup[i0]]_nbssgroup
         donnee.plus_T
         vectgroup.reste[ctr]_vectgroup[i0]
         vectgroup_vectgroup[-i0]
         ssgroup.plus_T
         ctr_ctr+1
      }
      else {donnee.plus_F}
   }
```

```
   if ((donnee.plus==F) | (taille(vectgroup)<=2)) {vectgroup_vectgroup.reste}
   list("ssgroup"=ssgroup,"ssgroup.plus"=ssgroup.plus,"vectgroup"=vectgroup)
}

# CREATION DE TOUS LES SOUS-GROUPES DANS TOUS LES GROUPES
suite_function(p,nbgroupe,groupe,n,donnees,ssgroup,cote.moyeg.vardif)
{
   nbssgroup_1
   for (i in 1:nbgroupe)
   {
      ssgroup.plus_T
      vectgroup_formgroup(i,groupe,n)
      j_1
      while (vectgroup[j]!=0)
      {
         ssgroup[vectgroup[j]]_nbssgroup
         j_j+1
      }
      nbssgroup_nbssgroup+1
      while((taille(vectgroup)>2) & (ssgroup.plus==T))
      {
         result_corps2(p,i,groupe,n,donnees,nbssgroup,vectgroup,ssgroup,
                       cote.moyeg.vardif)
         ssgroup_result$ssgroup
         ssgroup.plus_result$ssgroup.plus
         vectgroup_result$vectgroup
         if (ssgroup.plus==T) {nbssgroup_nbssgroup+1}
      }
   }
   list("ssgroup"=ssgroup,"nbssgroup"=nbssgroup)
}

# CALCUL DE LA TRACE DE LA VARIANCE DE TOUS LES SSGROUPES => VECTEUR VARIANCE
formvar_function(p,ssgroup, donnees, nbssgroup,n)
{
   formvar_rep.int(0,nbssgroup)
   for (i in 1:nbssgroup)
   {
      vect.ssgroup_formgroup(i,ssgroup,n)
      formvar[i]_trace.mat(cal.var(p,vect.ssgroup,donnees),p)
   }
   return(formvar)
}


# DETERMINE SI 2 SSGROUP I ET J APPARTIENNENT AU MEME GROUPE
```

```
memegr_function(k,l,mat.resume)
{
   i_1
   j_1
   while(mat.resume[3,i]!=k)
   {
      i_i+1
   }
   while(mat.resume[3,j]!=l)
   {
      j_j+1
   }
   numgrk_mat.resume[2,i]
   numgrl_mat.resume[2,j]
   if (numgrk==numgrl) {result_T} else {result_F}
   return(result)
}

# CALCUL LA MATRICE DE DISTANCE POUR LES TRACES DE VARIANCES
distvar_function(variance,n,mat.resume)
{
   M_matrix(0,n,n)
   for (i in 1:(n-1))
   {
      for (j in (i+1):n)
      {
         if ( is.na(variance[i])==F & is.na(variance[j])==F)
         {
            if((variance[j]!=0) & (variance[i]!=0) &(memegr(i,j,mat.resume)==F))
            {
               M[i,j]_d(variance[i],variance[j])
            }
         }
         if ((is.na(variance[i])!=F) | (is.na(variance[j])!=F) |
             (variance[j]==0) | (variance[i]==0) | (memegr(i,j,mat.resume)==T))
         {
            M[i,j]_ NA
         }
      }
   }
    return(M)
 }

# TROUVE LES 2 OBSERVATIONS APPARTENANT A DEUX GROUPES DIFFERENTS ET DONT
# LA DISTANCE EN TERME DE VARIANCE EST LA PLUS PETITE.
# TESTE SI ON PEUT REGROUPER CES DEUX OBSERVATIONS
```

```
corps3_function(p,surgroup,n,variance,ssgroup,vect.surgroup,
                nb.surgroup,M,donnees,arret,bic,nbssgroup,cote.moydif.vareg)
{
   vect.surgroup_rep.int(0,n)
   print("je passe ici")
   i0_trouvmin(M,nbssgroup)$i0
   j0_trouvmin(M,nbssgroup)$j0
   v_formgroup(i0,ssgroup,n)
   S1_cal.var(p,v,donnees)
   v_formgroup(j0,ssgroup,n)
   S2_cal.var(p,v,donnees)
   vect1.ssgroup_formgroup(i0,ssgroup,n)
   x1_calmoy(p,0,vect1.ssgroup)
   x1_t(t(x1))
   n1_taille(vect1.ssgroup)
   vect2.ssgroup_formgroup(j0,ssgroup,n)
   x2_calmoy(p,0,vect2.ssgroup)
   x2_t(t(x2))
   n2_taille(vect2.ssgroup)
   cote_cotemult.moydif(p,x1,x2, n1, n2, S1,S2)$cote
   if (cote>cote.moydif.vareg)
   {
         a_min(surgroup[vect1.ssgroup[1]],surgroup[vect2.ssgroup[1]])
         b_surgroup[vect1.ssgroup[1]]
         e_surgroup[vect2.ssgroup[1]]
         for (i in 1:n)
         {
            if (surgroup[i]==e | surgroup[i]==b)
            {
               surgroup[i]_a
            }
         }
         vect.surgroup_formgroup(a,surgroup,n)
         variance[i0]_trace.mat((S1+S2),p)
         variance[j0]_NA
         M[j0,]_NA
         M[i0,]_NA
         M[,j0]_NA
         M[,i0]_NA
         nb.surgroup_nb.surgroup-1
   }
   bicnouv_-2*log(max(cote,cote.moydif.vareg))+log(n)*nb.surgroup
   if(bicnouv<bic) {bic_bicnouv} else {arret_T}
   list("vect.surgroup"=vect.surgroup,"surgroup"=surgroup,
        "variance"=variance,"M"=M,"arret"=arret,"nb.surgroup"=nb.surgroup)
```

```
}

# CREE DES GROUPES DE MOYENNE DIFFERENTES MAIS DE VARIANCE EGALE
pro2_function(p,surgroup,n,variance,ssgroup,M, donnees,arret,nbssgroup,
              nb.surgroup,mat.resume,cote.moydif.vardif)
{
   bic_100
   arret_F
   group.plus_T
   nb.surgroup_nbssgroup
   vect.surgroup_rep.int(0,n)
   vect.surgroup[1]_1
   while ((group.plus==T) & (mat.vide(M,nbssgroup)==F) & (arret==F))
   {
      M_distvar(variance,nbssgroup,mat.resume)
      result_corps3(p,surgroup,n,variance,ssgroup,vect.surgroup,
                    nb.surgroup,M,donnees,arret,bic,nbssgroup,cote.moydif.vareg)
      vect.surgroup_result$vect.surgroup
      surgroup_result$surgroup
      variance_result$variance
      M_result$M
      nb.surgroup_result$nb.surgroup
      arret_result$arret
      if (vect.surgroup[1]!=0) {group.plus_T} else {group.plus_F}
      while ((vect.surgroup[1]!=0) & (mat.vide(M,nbgroupe)==F) & (arret==F))
      {
         result_corps3(p,surgroup,n,variance,ssgroup,vect.surgroup,
                       nb.surgroup,M,donnees,arret,bic,nbssgroup,cote.moydif.vareg)
         vect.surgroup_result$vect.surgroup
         surgroup_result$surgroup
         variance_result$variance
         M_result$M
         nb.surgroup_result$nb.surgroup
         arret_result$arret
      M_distvar(variance,nbssgroup,mat.resume)
      }
   }
   list("vect.surgroup"=vect.surgroup,"surgroup"=surgroup,
        "variance"=variance,"M"=M,"nb.surgroup"=nb.surgroup,"arret"=arret)
}

# CREATION DU JEU DE DONNEES MULTIVARIE
A_iris
Y_matrix(0,150,4)
Y[1:50,]_A[,,1]
Y[51:100,]_A[,,2]
```

```
Y[101:150,]_A[,,3]
Y_t(Y)
Y_Y[,-c(11:50,61:100,111:150)]
donnees_Y
X_donnees

n_150
p_4

cote.moyeg_1
cote.moyeg.vardif_10
cote.moydif.vareg_1

n0_0.00001

# CREATION DES GROUPES (voir section 3.1)
groupe_1:n
poids_rep.int(1,n)
vectgroup_rep.int(0,n)
arret_F
M_Dist(X,n)
result_pro(p,n,X,groupe,poids,vectgroup,M, n1, n2, donnees,arret,cote.moyeg)
groupe_result$groupe
poids_result$poids
nbgroupe_result$nbgroup
groupe_transgr(groupe,n)


# CREATION DES SOUS-GROUPES (voir section 3.1)
ssgroup_rep.int(1,n)
result_suite(p,nbgroupe,groupe,n,donnees,ssgroup,cote.moyeg.vardif)
ssgroup_result$ssgroup
nbssgroup_result$nbssgroup-1
ssgroup_transgr(ssgroup,n)

# RESUME DE CE QUI A ETE CREE
mat.resume_matrix(0,4,n)
mat.resume[1,]_1:n
mat.resume[2,]_groupe
mat.resume[3,]_ssgroup

# CREATION DES "SURGROUPES" (voir section 3.1)
surgroup_ssgroup
variance_formvar(p,ssgroup, donnees, nbssgroup,n)
M_distvar(variance,nbssgroup,mat.resume)
arret_F
```

```
result_pro2(p,surgroup,n,variance,ssgroup,M,donnees,arret,nbssgroup,nb.surgroup,
            mat.resume,cote.moydif.vareg )
nb.surgroup_result$nb.surgroup
variance_result$variance
surgroup_result$surgroup
surgroup_transgr(surgroup,n)
mat.resume[4,]_surgroup
dimnames(mat.resume)_list(c("observations","moy egale","moy eg et var diff",
                            "moy diff et var eg"),NULL)
```

Annexe B

ANNEXE B : MÉTHODES D'OPTIMISATION

```
##########################################
# PROGRAMME DES METHODES D'OPTIMISTATION #
##########################################



# CALCUL LA TRACE D'UNE MATRICE
trace.mat_function(A,n)
{
   trace.mat_0
   for (i in 1:n)
   {
      trace.mat_trace.mat+A[i,i]
   }
   return(trace.mat)
}

# DETERMINE QUELLES SONT LES OBSERVATIONS QUI FORMENT LE GROUPE K
formgroup_function(k,groupe,n)
{
   j_1
   vectgroup_rep.int(0,n)
   for (i in 1:n)
   {
      if (groupe[i]==k)
      {
         vectgroup[j]_i
         j_j+1
      }
   }
   return (vectgroup)
}

# DONNE L'INDICE MINIMUM ENTRE 3 VALEURS
```

```
ind.mini_function(a,b,c)
{
   m_min(a,b,c)
   if(m==a){k_1}
   else if(m==b){k_2}
   else {k_3}
   return(k)
}

# CALCUL LA MATRICE W POUR UNE PARTITION DONNEE
calW2_function(p,groupe,n,donnees)
{
   W_0
   for ( i in 1:3)
   {
       vectgroup_formgroup(i,groupe,n)
       if(vectgroup[1]!=0)
       {
          xb_rep.int(0,p)
          m_taille(vectgroup)
          for (i in 1:m)
          {
             xb_xb+donnees[,vectgroup[i]]
          }
          xb_xb/m
          xb_t(t(xb))
          S_matrix(0,p,p)
          for (i in 1:m)
          {
             val_donnees[,vectgroup[i]]
             val_t(t(val))
             S_S+(val-xb)%*%t(val-xb)
          }
          W_W+trace.mat(S,p)
       }
    }
    return(W)
}


# PROGRAMME PRINCIPAL
prog_function(groupe,n,p,donnees)
{
   for (i in 1:n)
   {
         groupe.anc_groupe
```

```
        groupe[i]_1
        W1_calW2(p,groupe,n,donnees)
        groupe_groupe.anc
        groupe[i]_2
        W2_calW2(p,groupe,n,donnees)
        groupe_groupe.anc
        groupe[i]_3
        W3_calW2(p,groupe,n,donnees)
        groupe_groupe.anc
        g_ind.mini(W1,W2,W3)
        groupe[i]_g
   }
   return(groupe)
}


A_iris
Y_matrix(0,150,4)
Y[1:50,]_A[,,1]
Y[51:100,]_A[,,2]
Y[101:150,]_A[,,3]
Y_t(Y)
donnees_Y

n_150
p_4

groupe_rep.int(0,n)

groupe[1]_1
groupe[51]_2
groupe[101]_3

groupe_prog(groupe,n,p,donnees)

groupe_prog.uni(groupe,n,donnees)
```

BIBLIOGRAPHIE

Akaike, H. (1978), A Bayesian analysis of the minimum AIC procedure, *Annals of the Institute of Statistical Mathematics*, **30**, 9-14.

Akaike, H. (1979), A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, *Biometrika*, **30**, 9-14.

Anderson, T.W. (1984), *An introduction to multivariate statistical analysis*, John Wiley & Sons Inc.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian theory*, John Wiley & Sons Inc, New York.

Binder, D.A. (1978), Bayesian cluster analysis, *Biometrika*, **65**, 31-38.

Box and Tiao (1992), *Bayesian inference in statistical analysis*, John Wiley & Sons Inc, New York.

Cormack, R.M. (1971), A review of classification, *Journal of the Royal Statistical Society*, (Série A), **134**, 321-367.

Everitt, B.S. (1993), *Cluster analysis*, John Wiley & Sons Inc, New York.

Geisser, S. et Cornfield, J. (1963), Posterior distributions for multivariate normal parameters, *Journal of the Royal Statistical Society*, (Série B), **25**, 368-376.

Gower, J.C. (1967), A comparaison of some methods of cluster analysis, *Biometrics*, **23**, 623-628.

Krzanowski, W.J. (1988), *Principles of multivariate analysis : A user's perspective*, Oxford University Press, Oxford.

Lance, G.N. et William, W.T. (1967), A general theory of classificatory sorting strategies : 1. Hierarchical systems, *Computer Journal*, **9**, 373-380.

Lorr, M. (1983), *Cluster analysis for social scientists*, Jossey-Bass Publishers, San Francisco.

Mariott, F.H.C. (1982), Optimization methods of cluster analysis, *Biometrika*, **69**, 417-421.

Robert, C. (1992), *L'analyse statistique Bayésienne*, Economica, Paris.

SAS/STAT (1989) User's Guide, Version 6, Edition 4, **1**, Cary, Caroline du Nord.

Singleton, R.C. et Karitz, W. (1965), *Minimum squared error clustering algorithm*, Stanford Research Institute, Stanford.

Sneath, P.H.A (1957), The application of computers to taxonomy. *Journal of Genetic and Microbiology*, **17**, 201-226.

Spath, H. (1980), *Cluster analysis algorithms*, Ellis Horwood Ltd, Chichester.

Symons, M.J. (1981), Clustering criteria and multivariate normal mixtures, *Biometrics*, **37**, 35-43.

Ward, Joe H. (1963), Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, **58**, 236-244.

Wishart, D. (1969), An algorithm for hierarchical classifications, *Biometrics*, **25**, 165-170.

Zellner, A. (1971), *An introduction to Bayesian inference in Econometrics*, John Wiley & Sons Inc, New York.

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *Bayesian inference and decision techniques*, P. Goel et A. Zellner (Eds.), Elsevier Publisher.