UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Spécialité : **Mathématiques Appliquées**

Présentée et soutenue publiquement

par

Isabel Brito

le 8 février 2002

Combinaison de modèles en analyse discriminante dans un contexte gaussien

---

Directeur de thèse : Gilles Celeux

JURY

Président :	Mme Anne Guérin-Dugué
Examinateurs :	M. António Silvestre (rapporteur)
	M. Jean-Pierre Nakache (rapporteur)
	Mme Helena Bacelar-Nicolau
	M. Olivier François
	M. Gilles Celeux (directeur de thèse)

Thèse préparée au sein du projet IS2 - INRIA RHÔNE ALPES

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Spécialité : **Mathématiques Appliquées**

Présentée et soutenue publiquement

par

Isabel Brito

le 8 février 2002

Combinaison de modèles en analyse discriminante dans un contexte gaussien

Directeur de thèse : Gilles Celeux

JURY

Président :	Mme Anne Guérin-Dugué
Examinateurs :	M. António Silvestre (rapporteur)
	M. Jean-Pierre Nakache (rapporteur)
	Mme Helena Bacelar-Nicolau
	M. Olivier François
	M. Gilles Celeux (directeur de thèse)

Thèse préparée au sein du projet IS2 - INRIA RHÔNE ALPES

À Lyes, Sammy, Inês et Miriam,

Avec tout mon amour.

Remerciements

C'est pour des raisons professionnelles que j'ai entrepris ma thèse. Si je n'étais pas profilée pour cette activité j'ai appris à aimer cet état d'esprit qui s'appelle "la recherche". Ce goût m'a été donné sans doute par l'encadrement dont j'ai bénéficié au sein de l'équipe IS2 de l'INRIA Rhône-Alpes. Pour que cette thèse aboutisse quelques années de travail se sont écoulés. Entre périodes de motivation et périodes de frustration j'ai réussi à mener ce bateau à bon port. Rien n'aurait été possible sans les appuis dont j'ai profité. Je voudrais ici remercier les personnes qui m'ont aidé avec ses contributions matérielles, ses encouragements ou tout simplement leur amitié.

En premier lieu, je tiens à exprimer ma profonde gratitude à Gilles Celeux. Il m'a suggéré ce travail, l'a dirigé et accompagné son déroulement. Ses critiques pertinentes m'ont permis d'améliorer constamment le travail. Tout cela est déjà assez pour son rôle de directeur de thèse. En plus je lui suis très reconnaissante pour sa grande disponibilité, son enthousiasme face au développement du travail, enfin sa sincère implication dans cette thèse. Aussi je tiens à lui remercier sa patience pour corriger mes fautes de d'orthographe.

J'ai été spécialement honorée que António Silvestre, professeur à l'Instituto Superior de Economia e Gestão de Lisbonne et Jean-Pierre Nakache, ingénieur de recherche à l'INSERM aient accepté d'être les rapporteurs de ma thèse. Je les remercie de l'avoir lue de façon aussi approfondie ainsi que pour les questions et critiques formulées de manière constructive.

Je remercie vivement Anne Guérin-Dugué, professeur à l'université Joseph

Fourier de Grenoble d'avoir accepté de présider mon jury.

J'adresse aussi de vifs remerciements à Helena Bacelar-Nicolau, professeur à la Faculté de Psicologia e de Ciências da Educação de Lisbonne et Olivier François, professeur à l'ENSIMAG de Grenoble pour avoir examiné ce travail.

Je voudrais remercier tous les membres de l'équipe IS2 que j'ai croisé. Ils sont un groupe formidable ; avec eux j'ai vécu une bonne expérience professionnelle et personnelle.

Enfin, je ne pourrais terminer sans dire mon merci ! à ma petite-grande famille (Lyes, Sammy, Inês et Miriam) qui a beaucoup souffert de ma manie "des x et des y".

Table des matières

Présentation de la thèse

La motivation de la thèse est la suivante. Dans un article de JASA en 1996, Bensmail et Celeux ont proposé une méthode de classification supervisée EDDA (Eigenvalue Decomposition Discriminant Analysis) se fondant sur la décomposition spectrale des matrices de variances des groupes à reconnaître. Cette méthode sélectionne un modèle d'analyse discriminante choisi parmi quatorze modèles différents. Ce faisant, on rejette treize modèles dont on a estimé les paramètres et qui peuvent apporter une information intéressante sur le problème d'analyse discriminante. Une autre approche que nous avons explorée dans cette thèse consiste à essayer de tirer parti des quatorze modèles potentiels en construisant une règle de décision les combinant d'une manière ou d'une autre plutôt que de sélectionner un unique modèle. Cet angle de vue peut parfois conduire à une règle de décision plus stable voire meilleure, ne serait ce que parce qu'un modèle n'est jamais parfaitement représentatif de la réalité. De plus, il peut s'avérer d'interprétation plus subtile que l'appel à un unique modèle.

Ainsi, partant de cet objectif d'évaluation des procédures de combinaison de modèles en analyse discriminante, le plan de la thèse est le suivant. Le chapitre 1 expose les prémisses, les objectifs ainsi que les procédures de l'analyse discriminante sur populations gaussiennes. En particulier les quatorze modèles de la méthode EDDA sont décrits.

La motivation pour la combinaison de modèles est d'exploiter les qualités individuelles des modèles par la construction d'un compromis entre modèles plutôt que de sélectionner un seul modèle. Dans le chapitre 2 nous retraçons les principaux axes de développement qu'a connu la combinaison de modèles dont les domaines d'utilisation privilégiés sont les séries chronologiques, l'apprentissage supervisé et les réseaux de neurones.

Le chapite 3 présente notre approche de la combinaison de modèles. Nous considérons deux types de combinaison de modèles : la combinaison en série et la combinaison hiérarchique. Pour le premier type, les modèles sont combinés de façon convexe en général sous une forme linéaire. Schématiquement, une combinaison linéaire de M modèles a la forme

$$p_{COM}(\mathbf{x}) = \sum_{m=1}^{M} p_m(\mathbf{x})\beta_m \tag{1}$$

où $p_{COM}(\mathbf{x})$ représente les probabilités par groupe à estimer par combinaison, $p_m(\mathbf{x})$ les probabilités par groupe obtenues par application du modèle m et β_m le poids attribué au modèle m dans la combinaison. Les $p_m(\mathbf{x})$ sont supposés connues, et il s'agit d'estimer les β_m. Par hypothèse ces pondérations sont indépendantes des observations $\mathbf{x}$ et les méthodes de combinaison diffèrent par leur stratégie d'obtention de "bons" coefficients β_m.

Dans la première section du chapitre 3 nous présentons trois méthodes classiques de combinaison en série qui reposent sur le critère des moindres carrés. Leur but commun est la minimisation de l'erreur de classement. Ces techniques se distinguent par l'expression du critère d'erreur à minimiser. Soit le vecteur $y_k,\ k = 1, \ldots, K$ qui pour chaque observation i vaut 1 si l'observation appartient au groupe k et 0 sinon. Nous envisageons les problématiques suivantes,

- une régression séparée pour chaque groupe :

  Les coefficients β sont obtenus par régression, l'erreur de régression est alors

$$e_{i,k} = y_{i,k} - \sum_{m=1}^{M} p_m^k(\mathbf{x}_i)\beta_m^k, \qquad k = 1, \ldots, K \tag{2}$$

  et les β_m^k sont obtenus par minimisation de $\sum_{i=1}^n e_{i,k}^2$;

- la vectorisation de la matrice des probabilités :

  L'erreur de régression est alors

$$e_{i,j}^* = y_{i,j} - \sum_{m=1}^{M}\sum_{k=1}^{K} p_m^k(\mathbf{x}_i)\beta_m^k, \qquad j = 1, \ldots, K \tag{3}$$

  et les β_m^k sont obtenus par minimisation de $\sum_{i=1}^n e_{i,j}^{*2}$;

- le comité de méthodes :

  cette technique utilise l'écart entre les probabilités par groupe des observations et les vecteurs binaires d'origine associés pour chaque modèle m

$$e_{i,k,m} = y_{i,k} - p_m^k(\mathbf{x}_i); \tag{4}$$

  partant de ces écarts, on considère

$$e_{u,v} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} e_{i,k,u}e_{i,k,v}, \qquad u, v = 1, \ldots, M; \tag{5}$$

  les coefficients β_m sont obtenus par minimisation de $\sum_{u=1}^M \sum_{v=1}^M \beta_u\beta_v e_{u,v}$, l'un des intérêts de cette méthode est de proposer un coefficient indépendant des groupes a priori, ce qui facilite leur interprétation.

Les techniques présentées dans la deuxième section du chapitre 3 visent à proposer un seul coefficient de pondération par modèle, indépendamment des groupes à discriminer. Elles s'appuient sur la vraisemblance des différents modèles. Nous considérons les méthodes suivantes, qui hormis la dernière, sont originales.

- les ratios de vraisemblance :

  Les β_m sont estimés proportionnellement aux ratios

  $$\frac{L_m(\theta; \mathbf{x})}{\sum_{u=1}^{M} L_u(\theta; \mathbf{x})}, \tag{6}$$

  où $L_m(\theta; \mathbf{x})$ représente la vraisemblance du modèle m.

- les ratios de vraisemblance pondérée ;

  La vraisemblance du modèle m est pondérée par le nombre de paramètres du modèle selon le critère d'information AIC. Les β_m sont estimés par

  $$\frac{L_m(\theta; \mathbf{x}) \exp^{-\nu_m}}{\sum_{u=1}^{M} L_u(\theta; \mathbf{x}) \exp^{-\nu_u}} \tag{7}$$

  où ν_m est le nombre de paramètres du modèle m.

- les ratios de vraisemblance intégrée ;

  Dans la logique de pondération bayésienne de modèles, nous considérons la vraisemblance intégrée. Dans l'impossibilité d'obtenir sa valeur exacte, on utilise le critère BIC pour approximer la vraisemblance intégrée. Les β_m sont alors estimés par

  $$\frac{L_m(\theta; \mathbf{x}) n^{-0.5\nu_m}}{\sum_{u=1}^{M} L_u(\theta; \mathbf{x}) n^{-0.5\nu_u}}. \tag{8}$$

- l'empilement :

  L'expression (1) peut être vue comme un modèle de mélange de mélanges. Dans cette perspective, estimer les coefficients β_m revient à estimer les proportions d'un mélange de densités connues $p_m^k(\mathbf{x}_i)$ pour $m = 1, \ldots, M$, ce qui se fait par l'algorithme EM.

La dernière section du chapitre 3 présente la combinaison hiérarchique. Ce type de combinaison est de tout autre nature et s'applique lorsque le nombre de groupes à discriminer est supérieur à deux. La méthodologie de cette section est entièrement nouvelle. Dans une combinaison du type hiérarchique les modèles sont emboîtés. Cela nous conduit à utiliser une méthode de couplage. Sur l'ensemble des groupes on construit des partitions en deux classes obtenues par minimisation du taux d'erreur estimé par validation croisée pour chaque modèle. Et, à chaque étape de la construction hiérarchique, on sélectionne le couple (modèle, partition) qui fournit le plus petit taux d'erreur estimé. Le résultat est représenté sous la forme d'un arbre où chaque branche est associé à un modèle et à une dichotomie entre des groupes à discriminer.

Dans le domaine de la combinaison de modèles, les résultats théoriques notamment asymtotiques sont inexistents et à vrai dire de peu d'utilité. Aussi une grande partie de notre travail a consisté à mener des expérimentations pour juger du bien-fondé des techniques de combinaison proposées, de leurs mérites comparés et de leurs principales caractéristiques par rapport au choix d'un unique modèle. Le chapitre 4 est le miroir de ces expérimentations où la performance des différents méthodes de combinaison est analysée. Nous avons ainsi expérimenté les techniques de combinaison envisagées dans le chapitre 3 pour combiner les quatorze modèles comparés dans la méthode EDDA.

Le fait que beaucoup de ces quatorze modèles fournissent des règles de décision analogues, limite l'intérêt de la combinaison de tous ces modèles. Aussi, dans un souci de mettre en évidence les gains possibles de la combinaison de modèles, nous avons également effectués des expérimentations en nous limitant aux trois modèles d'esprit le plus différent parmi les quatorze, à savoir l'analyse discriminante quadratique, l'analyse discriminante linéaire et le classement à la moyenne la plus proche. De plus dans un même ordre d'idée, nous avons mis au point une stratégie qui, préalablement à la combinaison, sélectionne les modèles les plus éloignés par une analyse des correspondances des probabilités par groupe fournies par chacun des modèles.

Enfin, dans un souci de bien explorer les possibilités pratiques de la combinaison de modèles en analyse discriminante, nous sommes sortis du contexte exclusivement gaussien. Ainsi, nous avons associé par combinaison, une méthode non paramétrique, d'inspiration locale, la méthode des k-plus proches voisins avec deux méthodes paramétriques, d'inspiration globale, l'analyse discriminante linéaire et le classement à la moyenne la plus proche.

Pour ces expérimentations, nous avons utilisé plusieurs jeux de données originaires du "Machine Learning Repository" de l'Université de Californie (Blake and Merz 1998), un jeu de données tiré du dépôt de l'Université d'Oxford (Venables and Ripley 1994) et un autre jeu de données publié par Habbema, Hermans, and van den Broek (1974). Voici une description abrégée de ces jeux de données. Chacun sera noté par un nom de référence.

bupa - Les données décrivent la présence ou l'absence de maladies du foie causées par une consommation excessive d'alcool. Pour 345 individus du sexe masculin ont été observées 7 variables.

crabes - Les observations en nombre de 200 portent sur 5 différentes mesures des carapaces de deux spécimens de crabes. Il faut distinguer entre 4 groupes : sexe et spécimen.

haberman - Le jeu de données contient 345 observations d'une étude sur la survie de femmes qui ont pratiqué une chirurgie du sein pour l'ablation d'un cancer. Il y a deux groupes : si la patiente est décédée dans une période de cinq ans apès l'opération ou si la patiente a survécu cinq ans ou plus.

haemo - Un échantillon de 75 des femmes caractérisées par 2 variables. Les femmes sont classifiées selon qu'elles sont porteuses ou pas d'hémophilie.

iris - Ce sont les bien connues données de Fisher : 4 variables décrivent 150 fleurs. Les fleurs appartiennent à 3 différents espèces.

thyroïde - Des tests de laboratoire sont pratiqués sur des individus pour prédire l'état de la thyroïde : normal, hyper ou hypo fonctionement. Nous disposons de 215 observations déterminées par 5 variables.

vin - Les données sont les résultats d'analyses chimiques effectuées sur des vins originaires de la même région en Italie mais produits par trois cultivateurs différents. Les analyses ont déterminé les quantités de 13 eléments trouvés dans chacun des trois types de vins. Le jeu de données rassemble 178 observations, réparties sur trois groupes et caractérisées par 13 variables.

De plus, nous menons une application de reconnaissance de tumeurs sur des données aimablement fournis par l'hôpital universitaire de Grenoble. L'objectif est la caractérisation tissulaire des tumeurs cérébrales *in vivo* chez l'homme. Le diagnostic repose sur des variables mesurées par imagerie spectroscopique.

Nous terminons cette thèse par un bilan résumant les conclusions que nous tirons de cette étude et dressons quelques perspectives. Nous avons analysé trois techniques de combinaison par régression. La régression séparée par groupe et la régression par la vectorisation de la matrice de probabilités estiment trop de coefficients[1]. Son interprétation devient compliquée et s'avère même impraticable. Toutes les autres techniques de combinaison en série analysées estiment un seul coefficient par modèle. En général les stratégies qui reposent sur la vraisemblance aboutissent à des meilleurs résultats. La combinaison hiérarchique s'est révélée particulièrement interessante. Dans le cas où le nombre de groupes

[1] Il y a un coefficient par modèle inclus dans la combinaison et par groupe.

est supérieur à 2 cette technique de combinaison semble prometteuse par sa simplicité. Finalement, nous croyons que la combinaison de modèles en analyse discriminante devra suivre plutôt soit une stratégie hiérarchique soit une stratégie en série de coefficients uniques par modèle fondée sur les vraisemblances.

Chapitre 1

Analyse Discriminante

L'esprit humain a besoin de catalogueur pour comprendre.

1.1 De quoi s'agit-il ?

Le domaine de l'Analyse Discriminante (AD) est la séparation statistique des groupes ou classes constituant une partition définie a priori. Toute une gradation de problèmes se pose en AD, qui va dès la simple description des différences des groupes jusqu'à la décision du groupe auquel doit appartenir un sujet d'origine inconnue. À titre de curiosité dans une perspective historique, l'une des premières utilisations consistantes du problème d'AD a été realisée par Fisher (1936). Il a proposé la fonction linéaire discriminante pour décrire les différences entre les groupes.

La désignation Analyse Discriminante est originaire du monde des statisticiens. Pourtant d'autres sciences se sont intéressées au même problème et ont essayé d'y répondre. C'est ainsi que "*supervised pattern recognition*" a vu le jour en informatique et sciences de l'ingénieur. Le rapport entre l'AD et d'autres branches de la connaissance ne s'arrête pas là. Plusieurs techniques ont émergé d'autres domaines et ont été assimilées. Comme exemple, il y a les réseaux de neurones qui viennent de l'Intelligence Artificielle. Une bonne monographie à ce propos est celle de Ripley (1996).

Pour aboutir à chacun des objectifs de l'AD la population est décrite par des caractéristiques mesurables ou variables qui vont alimenter les méthodes. Du point de vue descriptif, l'AD cherche à découvrir les variables qui expliquent au mieux les différences entre les groupes et à déterminer si elles sont significatives. Du point de vue décisionnel, l'AD construit une règle d'affectation des sujets à l'un des groupes qui puisse être appliquée dans le futur. Cette règle est basée sur les variables et on espère qu'elle conduise à des erreurs de classement le plus faibles possible. Parfois les deux perspectives se confondent dans un même problème. Ainsi, l'AD cherche des méthodes qui visent à établir des règles de décision dont le pouvoir explicatif soit grand.

Nous poursuivons dans ce travail la perspective décisionnelle. On désire, alors, construire une bonne règle de classement et pour y aboutir la théorie de la décision statistique (Berger 1980) est un outil indispensable.

1.2 Cadre probabiliste

Voici le cadre probabiliste dans lequel est fondée l'AD. Nous voulons classer un objet d'après son appartenance à un de K groupes $G_1, ..., G_K$ connus a priori. Chaque objet est décrit par un vecteur $\mathbf{x}$ de d variables.

Ce vecteur $\mathbf{x}$ est considéré comme une réalisation du vecteur aléatoire $\mathbf{X}$ de dimension d. $\mathbf{X}$ est constitué par les variables aléatoires explicatives $\{X_1, ..., X_d\}$. On traitera ici le cas des variables quantitatives.

Le vecteur $\mathbf{X}$ est associé à la variable aléatoire Z qui traduit le groupe d'appartenance des sujets et qui prend, ainsi ses valeurs dans l'ensemble $\{1, ..., K\}$. La réalisation z est souvent remplacée par le vecteur indicateur binaire $\mathbf{y}$ de dimension K. Le $k$ième élément de $\mathbf{y}$, y_k vaut 1 si le sujet appartient au groupe G_k, $(k = 1, \ \ldots, \ K)$ et 0 sinon.

On considère que la variable aléatoire Z est munie d'une loi de probabilité $\{\pi_1, \ \pi_2, \ ..., \ \pi_K\}$ où la quantité

$$\pi_k = P(Z = k), \quad \pi_k \geq 0 \tag{1.1}$$

représente la probabilité a priori du groupe G_k. Cela équivaut à déclarer que la variable aléatoire $\mathbf{Y}$ suit une loi multinomial avec

$$P(\mathbf{Y} = \mathbf{y}) = \pi_1^{y_1} \pi_2^{y_2} ... \pi_K^{y_K}. \tag{1.2}$$

Étant donné que l'ensemble $\mathcal{G}$ des groupes, $\mathcal{G} = \{G_1, ..., G_K\}$, est une partition, alors,

$$\sum_{i=1}^{K} \pi_k = 1. \tag{1.3}$$

On suppose que pour les sujets appartenant au groupe G_k le vecteur $\mathbf{X}$ suit une loi de densité

$$f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} | Z = k) \tag{1.4}$$

par rapport à une mesure de référence. La densité marginale de $\mathbf{X}$ s'écrit alors

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}). \quad (1.5)$$

Cette densité est celle d'un mélange des densités conditionnelles par groupes $f_k(\mathbf{x})$ dont les proportions sont les probabilités a priori d'apparition des groupes π_k.

Avec $\mathbf{x}$ ayant été observé, la probabilité a posteriori qu'un sujet appartienne au groupe G_k est

$$p_k(\mathbf{x}) = P(Z = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{l=1}^{K} \pi_l f_l(\mathbf{x})}. \quad (1.6)$$

1.3 La règle de décision

La décision consiste à affecter un sujet dont on ne connaît pas le groupe auquel il appartient à l'un des K groupes prédéfinis. Les règles de décision à construire doivent produire un minimum d'erreurs lorsqu'elles seront appliquées dans le futur. Soit $\delta(\mathbf{x})$ une règle de décision, c'est une fonction de $\mathcal{X}$ dans $\{1, ..., K\}$ où $\delta(\mathbf{x}) = k$ signifie que les sujets associés au vecteur $\mathbf{x}$ sont affectés au groupe G_k. L'espace $\mathcal{X}$ est celui où le vecteur $\mathbf{X}$ prend ses valeurs, $\mathcal{X} \subset \Re^d$. La règle conduit à une partition de $\mathcal{X}$ en K régions $R_1, ..., R_K$. Si $\delta(\mathbf{x}) = k$, alors $\mathbf{x}$ se situe dans région R_k.

La probabilité qu'un sujet appartenant au groupe G_k soit affecté au groupe G_l $(k, l = 1, ..., K)$ est

$$P(\delta(\mathbf{x}) = l|Z = k) = \int_{R_l} f_k(\mathbf{x}) d\mathbf{x}. \quad (1.7)$$

La probabilité qu'un sujet appartenant au groupe G_k soit mal affecté peut alors s'écrire

$$P(\delta(\mathbf{x}) \neq k|Z = k) = \int_{\Re^d - R_k} f_k(\mathbf{x})d\mathbf{x}. \tag{1.8}$$

Pour mesurer la qualité de la règle de décision il faut introduire la matrice des coûts de mauvais classement $C(l|k)$. Elle représente les coûts de classement dans le groupe G_l d'un individu du groupe G_k,

$$C(l,k) = C(\delta(\mathbf{x}) = l, Z = k). \tag{1.9}$$

On pose $C(l,k) = 0$ si $l = k$ et dans le cas où $l \neq k$, si l'on suppose l'égalité des coûts, on pose $C(l,k) = 1$ afin de simplifier les calculs. Ainsi tous les mauvais classements ont le même coût.

La fonction de risque d'une règle de décision $R(\delta(\mathbf{x})|Z)$ est définie comme l'espérance de son coût sachant le vrai groupe

$$R(\delta(\mathbf{x}) = l|Z = k) = \sum_{l=1}^{K} C(\delta(\mathbf{x}) = l, Z = k) \int_{R_l} f_k(\mathbf{x})d\mathbf{x}. \tag{1.10}$$

Le risque moyen d'une fonction de décision $R(\delta(\mathbf{x}))$, aussi appelé risque de Bayes, est l'espérance de son coût

$$R(\delta(\mathbf{x})) = \int_{\mathcal{X}} [\sum_{k=1}^{K} C(\delta(\mathbf{x}), k) P(Z = k|\mathbf{X} = \mathbf{x})] f(\mathbf{x})d\mathbf{x}. \tag{1.11}$$

Notre objectif est naturellement d'obtenir la règle de décision $\delta^*(\mathbf{x})$ qui minimise le risque de Bayes. On dit alors que $\delta^*(\mathbf{x})$ est la règle de Bayes. En observant l'expression (1.11) on remarque que pour l'obtenir il suffit de minimiser l'espérance conditionnelle

$$\sum_{k=1}^{K} C(\delta(\mathbf{x}), k) P(Z = k|\mathbf{X} = \mathbf{x}) \tag{1.12}$$

par rapport à $\delta(\mathbf{x})$ pour chaque $\mathbf{x}$. Le minimum est atteint pour le choix de $\delta^*(\mathbf{x}) = l$ tel que la classe G_l est celle qui a la probabilité a posteriori $p_k(\mathbf{x})$ la plus élevée,

$$\delta^*(\mathbf{x}) = \arg\max_k \; p_k(\mathbf{x}). \tag{1.13}$$

1.4 Estimation des probabilités a priori et des densités conditionnelles par groupe

Pour établir la règle de Bayes $\delta^*(\mathbf{x})$ il faut connaître les probabilités a priori π_k et les densités conditionnelles par groupe $f_k(\mathbf{x})$[1]. Ces quantités ne sont pas connues et doivent être estimés de façon à ce que le risque de Bayes empirique obtenu soit le plus proche possible du risque théorique $R(\delta)$. La règle de décision ainsi obtenu s'appelle règle "*plug-in*". On peut aussi estimer les probabilités a posteriori $p_k(\mathbf{x})$ directement comme dans le modèle logistique mais celui-ci ne sera pas exploité dans le présent texte. Nous poursuivons la voie indirecte de la règle "*plug-in*" sur les densités conditionnelles par groupe.

Le point de départ pour l'estimation des π_k et des $f_k(\mathbf{x})$ est l'existence d'un échantillon $\mathbf{t}$, dit d'apprentissage, de n observations de la population décrite par $\mathbf{X}$ et dont l'appartenance à l'un des K groupes est connu. Ainsi chaque sujet de l'échantillon est noté $\mathbf{t}_i = (\mathbf{x}_i, z_i)$, $(i = 1, \ldots, n)$.

L'estimation de π_k dépend du schéma d'échantillonnage utilisé. Les deux schémas possibles sont le schéma de mélange et le schéma rétrospectif à K échantil-

[1] À cause de (1.13) et (1.6).

lons. Celui que nous privilégions dans ce travail est le schéma de mélange pour lequel les observations sont tirés aléatoirement sans distinction de groupes. De cette façon la taille n_k du sous-échantillon issu de G_k est aléatoire. On suppose que les données sont indépendantes et identiquement distribuées. Les estimateurs du maximum de vraisemblance des π_k sont

$$\hat{\pi}_k = \frac{n_k}{n}. \tag{1.14}$$

Dans le cas du schéma rétrospectif, l'échantillon d'apprentissage $\mathbf{t}$ est le regroupement de K échantillons indépendants de taille fixé n_k et de loi $f_k(\mathbf{x})$. Alors, les n_k ne sont pas aléatoires et ne peuvent servir à estimer les probabilités π_k. Dans ce cas il est usuel considérer toutes les classes a priori à parts égales en faisant l'hypothèse que $\hat{\pi}_k = \frac{1}{K}$.

Fondamentalement, l'estimation statistique d'une règle de décision consiste à estimer les densités conditionnelles par groupe $f_k(\mathbf{x})$. Les différentes méthodes d'AD se distinguent par la modélisation faite sur ces densités et les techniques d'estimation utilisées. Pour estimer les $f_k(\mathbf{x})$ il y a deux approches. L'une, l'approche paramétrique qui suppose les $f_k(\mathbf{x})$ appartenant à une famille paramétré de densités

$$\{f_k(\mathbf{x}, \theta_k) : \theta_k \in \Theta_k\} \tag{1.15}$$

où θ_k est un vecteur de paramètres inconnu qui appartient à l'espace de paramètre Θ_k. Les modèles EDDA qui seront présentés dans la section 1.6 rentrent dans ce genre de méthodes.

L'autre approche ne fait pas d'hypothèse paramétrique sur les densités par groupe. Les deux principales techniques non paramétriques sont la méthode des noyaux et la méthode des k plus proches voisins. Les deux méthodes sont bien décrites par exemple dans McLachlan (1992).

Souvent on suppose que les K densités conditionnelles $f_k(\mathbf{x})$ appartiennent à la même famille paramétrée. Le plus courant est de supposer des lois normales d-variés avec $\theta_k = (\mu_k, \Sigma_k)$ où μ_k est le vecteur moyen de la population appartenant au groupe G_k et Σ_k la matrice de variance-covariance de la même population.

L'AD linéaire et l'AD quadratique sont les deux cas classiques de discrimination qui se placent sous l'hypothèse de densités par groupes normales. Dans le cas linéaire on suppose que les matrices variance-covariance des groupes sont identiques ($\Sigma_1 = \Sigma_2 \ \ldots \ = \Sigma_K = \Sigma$) alors que dans le cas quadratique on permet qu'elles soient différentes.

À partir des données réunis dans l'échantillon d'apprentissage $\mathbf{t}$, θ_k est obtenu et remplacé par son estimateur $\hat{\theta}_k$ du maximum de vraisemblance.

1.5 Le taux d'erreur

La règle de décision δ est construite de façon à produire un minimum d'erreurs lorsqu'elle sera appliquée sur de nouveaux cas. Une étape importante en AD est celle de l'évaluation de la performance de la règle de décision.

Supposons que la règle de décision dépend du vecteur de paramètres θ, que les coûts de mauvais classement sont tous égaux à un et que les π_k sont égaux. Le risque de la règle de décision est dans ce cadre son taux inconditionnel d'erreur, noté $T(\delta)$. Le vecteur θ est inconnu et est estimé égal à $\hat{\theta}$. On distingue trois types de taux d'erreur.

- le taux d'erreur optimal - $T(\delta(\theta))$ - associé à la règle optimale $\delta(\theta)$ on le calcule si le vecteur θ est connu.
- le taux d'erreur exact - $T(\delta(\hat{\theta}))$ - associé à la règle de classement estimée $\delta(\hat{\theta}))$ on le calcule par subtitution de θ par $\hat{\theta}$ dans la règle optimale.
- le taux d'erreur estimé - $\hat{T}(\delta(\hat{\theta}))$ - l'estimateur du vrai taux d'erreur $T(\delta(\hat{\theta}))$.

Le taux d'erreur optimal peut rarement être calculé et le taux d'erreur exact n'est pas calculable car $\delta(\hat{\theta})$ ne peut être appliquée indéfiniment. Ainsi, mesurer la performance de $\delta(\hat{\theta})$ consiste à proposer un estimateur $\hat{T}(\delta(\hat{\theta}))$.

Lorsque la règle de décision a été construite sous des hypothèses paramétriques explicites, il est possible de calculer $\hat{T}(\delta(\hat{\theta}))$. Pourtant cette pratique est peu utilisée. Ainsi les principales procédures d'estimation du taux d'erreur sont l'estimation par resubstitution, par échantillon test, par validation croisée, par jackknife et par bootstrap (Celeux and Nakache 1994).

Dans cette thèse nous estimons le taux d'erreur par la validation croisée. Ce choix a deux motivations. D'une part la technique de la validation croisée est assez précise, plus que la resubstitution ou même que l'échantillon test, et est moins lourde que le bootstrap. D'autre part, elle est la technique adopté dans le cas des modèles EDDA qui constituent le cadre de la thèse.

La procédure de validation croisée consiste à retirer chaque observation $\mathbf{t}_i$, $(i = 1, \ldots, n)$ de l'échantillon d'aprentissage $\mathbf{t}$ et à construire la règle de décision δ_{-i} sur la base de $\mathbf{t}$ privé de $\mathbf{t}_i$. Finalement le vecteur $\mathbf{x}_i$ est classé sur la base de la règle δ_{-i}. Soit

$$T_i = \begin{cases} 1 & \text{si } \mathbf{x}_i \text{ n'est pas classé dans son groupe d'origine } z_i \\ 0 & \text{sinon} \end{cases}$$

L'estimation par validation croisée du taux d'erreur est,

$$\hat{T}(\delta) = \frac{1}{n} \sum_{i=1}^{n} T_i. \tag{1.16}$$

Une généralisation de la validation croisée consiste à découper l'échantillon $\mathbf{t}$ de taille n en V parties de même taille $\mathbf{t}^1, \ldots, \mathbf{t}^V$. Ensuite on construit la règle de décision δ_{-v}, $(v = 1, \ldots, V)$ sur la base de l'échantillon $\mathbf{t}$ privé de la partie $\mathbf{t}^v$ et les observations de la partie $\mathbf{t}^v$ sont affectées par la règle δ_{-v}. Dans ce cas l'estimation du taux d'erreur est,

$$\hat{T}(\delta) = \frac{1}{V} \sum_{v=1}^{V} T_v, \tag{1.17}$$

où T_v représente le nombre de points de $\mathbf{t}^v$ mal classés par la règle δ_{-v}.

Pour distinguer la validation croisée de sa généralisation on les nomme validation croisée par unité ou standard (*leave-one-out*) pour la première et validation par blocs pour la seconde. Dorénavant l'estimation du taux d'erreur $\hat{T}(\delta)$ sera appelé taux d'erreur de classement (TEC).

1.6 Les modèles EDDA

EDDA vient de l'anglais “Eigenvalue Decomposition Discriminante Analysis”. Ces modèles se présentent comme une panoplie autour des deux cas classiques : l'AD linéaire et l'AD quadratique, en régularisant les matrices de variance-covariance des groupes Σ_k.

Ils ont émergé comme une approche alternative à l'analyse discriminante régularisée de Friedman (1989). Étant proposés par Bensmail and Celeux (1996), ils sont fondées sur la reparamétrisation de la matrice de variance-covariance du groupe G_k, Σ_k, au moyen de sa décomposition spectrale. Ainsi

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k \tag{1.18}$$

où $\lambda_k = \mid \Sigma_k \mid^{1/d}$, $\mathbf{D}_k$ est la matrice des vecteurs propres de Σ_k et $\mathbf{A}_k$ est une matrice diagonale avec $\mid \mathbf{A}_k \mid = 1$ qui contient sur la diagonale les valeurs propres normalisés de Σ_k par ordre décroissant.

Les atouts de ces modèles sont ses bonnes performances et aussi son interprétabilité. La reparamétrisation (1.18) donne lieu à des paramètres qui ont un sens géométrique. Ainsi, λ_k représente le volume des contours de la densité de la classe G_k, $\mathbf{A}_k$ sa forme et $\mathbf{D}_k$ son orientation.

En permettant que quelques unes de ces quantités varient selon les groupes il est possible d'avoir des modèles de discrimination facilement interprétables. Différentes hypothèses sur les paramètres font émerger 14 modèles classés en trois types. Pour le premier type, nous considérons d'abord les hypothèses générales sur les paramètres λ_k, $\mathbf{D}_k$ et $\mathbf{A}_k$. Cela nous conduit à huit modèles à savoir,

1er type $[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$, $[\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}']$, $[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$, $[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$
$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$, $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$, $[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$, $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$.

La présence de l'indice k veut dire que le paramètre en question n'est pas le même pour tous les groupes et son absence indique qu'il est le même pour tous les groupes. Ainsi, par exemple, le modèle $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ signifie qu'on considère que les densités des groupes ont le même volume, la même forme et des orientations différentes. Le modèle $[\lambda \mathbf{DAD}']$ est le modèle linéaire et $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$ est le modèle quadratique.

Pour le deuxième type de modèles l'hypothèse est que les matrices de variance-covariance Σ_k sont diagonales. Dans ce cas, les matrices d'orientation sont des matrices de permutation, et les changements d'orientation ne sont pas intéressants. C'est pourquoi la simplification $\Sigma_k = \lambda_k \mathbf{B}_k$ est possible, $\mathbf{B}_k$ étant une matrice diagonale dont $|\mathbf{B}_k| = 1$. Quatre modèles sont construits,

2ème type $\quad [\lambda \mathbf{B}],\ [\lambda_k \mathbf{B}],\ [\lambda \mathbf{B}_k],\ [\lambda_k \mathbf{B}_k]$

qui indiquent des variations de volume ou de forme.

Finalement le troisième type considère des modèles avec des formes sphériques. La matrice $\mathbf{A} = \mathbf{I}$ ce qui donne lieu à deux modèles,

3ème type $\quad [\lambda \mathbf{I}],\ [\lambda_k \mathbf{I}]$

indiquant des variations de volume. Au total pour les trois types on aura quatorze modèles.

Pour chaque modèle les vecteurs moyens des groupes μ_k sont estimés par le maximum de vraisemblance. Il s'agit de la moyenne empirique de l'échantillon du groupe

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{\sum^{i|z_{i=k}} \mathbf{x}_i}{n_k}. \tag{1.19}$$

Les estimateurs du maximum de vraisemblance pour les matrices de variance-covariance des groupe Σ_k sont obtenues pour le modèle en question selon des formules ou algorithmes décrits dans Bensmail and Celeux (1996).

Le meilleur modèle, que l'on choisira parmi les quatorze, est celui dont le taux d'erreur de classement (TEC) obtenu par validation croisée est minimum. Si plusieurs modèles ont le même TEC le modèle préféré est le modèle le plus parcimonieux, celui qui a le moins de paramètres[2]. Si des modèles différents ont le même nombre de paramètres la règle d'ordre est : un modèle sphérique est préféré à un modèle diagonal lequel à son tour est préféré à un modèle non diagonal ; un modèle avec volumes différents est préféré à un modèle avec formes différentes et celui-ci est préféré à un modèle avec orientations différentes.

1.7 La méthode des k plus proches voisins

Une voie pour l'estimation non paramétrique des densités des classes est d'utiliser la règle des k plus proches voisins (k-ppv). La probabilité qu'un point de G_k tombe dans un petit voisinage de volume M d'un point $\mathbf{x}$ est aproximativement égale à $f_k(\mathbf{x})M$. Si b_k est le nombre de points de G_k tombant dans le petit voisinage, une estimation naturelle de $f_k(\mathbf{x})$ est

$$\hat{f}_k(\mathbf{x}) = \frac{b_k}{n_k M}. \tag{1.20}$$

A partir de là on fixe le nombre b de voisins d'un point auquel cas le voisinage dépendra des données et variera suivant les zones de concentration de l'échantillon. Toutefois, l'estimateur des k-ppv de $f_k(\mathbf{x})$ n'est pas une densité car $\int_{\mathcal{X}} \hat{f}_k(\mathbf{x})d\mathbf{x} = \infty$.

[2]L'estimation de μ_k concerne $k.d$ paramètres et l'estimation de Σ_k concerne un nombre variable de paramètres. Pour le modèle le plus complexe, le modèle quadratique, il y a $(\frac{d.(d-1)}{2} + d).k$ paramètres.

Si l'on suppose que $\hat{\pi}_k = n_k/n$, la probabilité a posteriori que $\mathbf{x}$ appartienne à G_k selon la règle des k-ppv est approximé par,

$$p_k(\mathbf{x}) \simeq \frac{f_k(\mathbf{x})\pi_k}{f(\mathbf{x})} = \frac{b_k}{n_k M}\frac{n_k}{n}/\frac{b}{nM} = \frac{b_k}{b}.$$

La règle discriminante est donc : on affecte tout point $\mathbf{x}$ à la classe majoritaire parmi ses k-ppv.

La qualité de la discrimination dépend fortement du nombre choisi b de voisins. Il est souvent nécessaire de faire varier ce nombre b pour obtenir les meilleurs résultats possibles. Une possibilité efficace est de choisir b par validation croisée.

Chapitre 2

La combinaison de modèles - État de l'art

Deux têtes ne raisonnent pas forcément
mieux que une seule mais
leurs jugements peuvent se compléter.

2.1 Les motivations pour une combinaison

Combiner pourquoi ?

Le but de l'analyse discriminante est d'accomplir la meilleure performance de classification pour le problème à résoudre. Cet objectif mène souvent au développement de plusieurs modèles de classification que sont ainsi mis en concurrence. Une évaluation expérimentale appelée validation justifie le choix de l'un des modèles comme solution du problème. Quel que soit le modèle choisi, il a été observé que les sujets mal classées par les différents modèles ne sont pas toujours les mêmes. De là la suggestion que les modèles rejetés offriraient des informations complémentaires sur l'ensemble à classer et ainsi pourraient être utilisés afin d'améliorer la performance du modèle choisi. À cela s'ajoute le fait que l'effort en programmation supplémentaire est réduit. Ainsi, il est certain que l'estimation des modèles rejetés n'aurait pas été du simple gaspillage.

Le recours à différents modèles peut apparaître en différentes circonstances. À savoir,

- Un analyste peut avoir accès à plusieurs modèles chacun développé dans un contexte différent et pour une représentation distincte du même problème. Comme exemple considérons l'identification des personnes par leur voix, leur visage et aussi par leur écriture.
- Parfois plus d'un échantillon d'apprentissage est disponible ; chacun est obtenu à un moment différent ou dans un environnement différent. Ces échantillons peuvent même utiliser des variables explicatives différentes.
- Des modèles basés sur les mêmes données peuvent différer non seulement dans leurs performances globales mais ils peuvent présenter de grandes différences locales. Chaque modèle peut, en effet, avoir une meilleure performance dans une région particulière de l'espace des variables explicatives.
- Les modèles de réseaux de neurones dépendent beaucoup de leur initialisation. Au lieu de sélectionner le meilleur et rejeter les autres, on peut combiner plusieurs réseaux.

Ces remarques ont motivé la récente recrue d'intérêt pour la combinaison de modèles bien que les origines de ce thème ne datent pas d'aujourd'hui. Dans bon nombre de cas la classification fondée sur un schéma de combinaison s'avère compétitif par rapport aux modèles uniques.

L'idée de la combinaison est de pouvoir ne pas compter seulement sur un modèle mais que le schéma de décision soit un consensus basé sur tous ou quelques-uns des modèles. En résumé il peut y avoir des ensembles de variables différents, des échantillons différents, des méthodes de discriminations différentes ou des modélisations différentes, le tout aboutissant à un ensemble de modèles dont les résultats peuvent être combinés. La discrimination obtenue de cette façon poursuit toujours le but d'augmenter la capacité prédictive du modèle formé par combinaison. Il se peut aussi que l'objectif soit d'avoir à la fois des modèles plus stables et plus robustes puisque ne dépendant pas des hypothèses particulières d'un seul modèle.

La motivation pour la combinaison de modèles est, en un mot, de tirer profit des bonnes qualités des modèles particuliers en sachant que l'effort supplémentaire en programmation est négligeable. Dans certains cas il est possible que le modèle combiné conduise a une règle finale moins transparente mais cela au profit d'une amélioration du taux d'erreur. Dans d'autres cas, au contraire, la règle discriminante gagne en interprétabilité.

Winckler (1989) a réfléchi aux fondements philosophiques de la combinaison de modèles. L'analyse discriminante a suivi la tradition de la statistique classique de modélisation dans l'esprit de recherche du vrai modèle. La combinaison de modèles semble aller contre cet esprit ; elle peut être pragmatique dans le court terme mais dans le long terme ne peut qu'empêcher la recherche du vrai modèle. Toutefois si la procédure de combinaison se traduit par une moyenne pondérée des modèles en compétition alors les poids attribués à chaque modèle sont interprétés comme étant la probabilité que le modèle soit le vrai. Ainsi la combinaison peut être placée dans une démarche classique.

Pourtant Winckler préfère un autre point de vue. Il croit que souvent le vrai modèle n'existe pas dans le domaine de la prévision. Le monde change tout le temps, c'est pourquoi la recherche du vrai modèle dans le long terme est une tâche sans espoir. Par conséquent les modèles doivent être adaptables. La combinaison de modèles agrège l'information afin de réduire l'incertitude ou augmenter la précision.

D'un point de vue bayésien la combinaison de modèles est perçue comme une méthode où le décideur a besoin d'utiliser au mieux les modèles disponibles.

La perspective de combinaison de modèles versus le choix du meilleur modèle si elle n'est pas nouvelle a néanmoins pris un fort élan récemment. Les méthodologies de combinaison proposées ont émergé, éparpillés dans plusieurs communautés de recherche, comme la statistique, les réseaux de neurones, l'apprentissage supervisé et encore les séries chronologiques, pour ne citer que les plus importantes. Ces schèmas de combinaison sont surtout heuristiques bien que quelques tentatives de fournir un cadre formel commun à quelques règles aient déjà surgi.

En statistique le thème de la combinaison d'estimateurs au détriment du choix du meilleur a bonne presse. Quelques références sont, par exemple, Rao and Subrathmaniam (1971), Efron and Morris (1973), Rubin and Weisberg (1975), Berger and Bock (1976) et Green and Strawderman (1991). Efron et Morris veulent estimer deux ensembles de paramètres et déterminent des situations où l'estimateur de James-Stein doit être appliqué sous une forme combinée. Rubin et Weisberg considèrent deux types de combinaison d'estimateurs indépendants non biaisés d'un certain paramètre. L'une est la combinaison linéaire d'estimateurs de variance minimale et l'autre est n'importe quelle combinaison dont la somme des poids, ceux-ci indépendants des estimateurs, soit un.

Aussi dans la communauté neuronale la modélisation entraîne souvent l'essai de plusieurs modèles mis en concurrence et celui qui démontre la meilleure performance est choisi tandis que les autres sont écartés. Parfois plusieurs petits réseaux sont construits et ensuite combinés au lieu de considérer un réseau plus grand. Le domaine des réseaux de neurones est intéressant par la manière dont le problème des minimums locaux y est traité. En effet les pondérations agissent sur l'espace des fonctions et non sur celui des paramètres, ainsi le problème des minimums locaux ne se pose pas quand on fait la combinaison de petits réseaux. La combinaison de réseaux de neurones, fondés sur différentes initialisations a été étudié par Hashem and Schmeiser (1995), Cho and Kim (1995a), Cho and Kim (1995b), Hansen and Salamon (1990), Krogh and Vedelsby (1995) et Rogova (1994).

En séries chronologiques la combinaison de modèles à but prévisionnel a été développée. Les modèles combinés sont appelés méthodes hybrides par opposition à méthodes singulières. Des références sont, par exemple, Granger and Ramanathan (1984), Clemen (1989), Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen, and Winckler (1982), Schnaars (1984), Sanders and Ritzman (1990), Winckler (1989) et Arinze, Kim, and Anandarajan (1997). Clemen nomme plusieurs travaux dans sa revue d'articles concernant la combinaison tirés des journaux ayant trait à la prévision, la psychologie, la statistique et la gestion.

L'idée de combinaison s'est généralisée dans maints domaines mais aussi divers qu'ils peuvent être, leur ressort commun est l'analyse de données. Nous décrions maintenant les techniques de combinaison qui ont surgi et qui peuvent être adaptées à l'analyse discriminante.

2.2 Les méthodologies de combinaison

Combiner comment ?

Si l'ensemble de modèles est déterminé la recherche se concentre sur la fonction de combinaison. Les techniques de combinaison de modèles proposées jusqu'à présent sont très diverses. Essentiellement nous les regroupons en deux types d'après les principes de combinaison utilisés. Ainsi, on considère l'approche différenciation de modèles et l'approche modèle unique.

Le mode d'usage des données échantillonnées déterminera laquelle des deux approches est utilisée. On suppose que les modèles composants de la combinaison sont tous fondés sur le même échantillon d'apprentissage mais correspondent formellement à des approches différentes tels que les réseaux de neurones, les voisins plus proches, etc., ou ils sont du même type mais diffèrent en complexité tels les modèles EDDA vus au chapitre précédent. Dans ce cas nous sommes en présence d'une combinaison du type différenciation de modèles.

L'autre mode de combinaison de modèles s'appuie sur des stratégies de reéchantillonnage comme le bootstrap ou le boosting. Chaque modèle est reproduit avec le même algorithme mais se base sur différents échantillons d'apprentissage. Il s'agit d'une combinaison du type modèle unique.

Dans la présente thèse on s'intéresse à la voie différenciation de modèles, en privilégiant les modèles gaussiens. Ainsi, on décrit cette approche, sans oublier de faire quelques références sur l'approche modèle unique.

Notations

Tout d'abord et pour fixer le cadre de l'étude, reprenons le scénario habituel de l'analyse discriminante. Ainsi, on dispose de l'échantillon d'apprentissage $\mathbf{t}$ de n observations. Chaque observation s'écrit $\mathbf{t}_i = (\mathbf{x}_i, z_i)$, $(i = 1, ..., n)$ où $\mathbf{x}_i$ décrit d mesures pour l'individu i et z_i indique son groupe d'appartenance $z_i \in \{1, \ldots, K\}$. Considérons M modèles qui servent à classer des observations d'origine inconnue. L'entrée des modèles est l'échantillon t et leur sortie est, soit une estimation des densités conditionnelles par groupe $f_m^k(\mathbf{x}_i)$, $(m = 1, \ldots, M, \ k = 1, \ldots, K.)$, ou des probabilités a posteriori $p_m^k(\mathbf{x}_i)$ ou du groupe d'appartenance des observations à classer $z_m(\mathbf{x}_i)$ (cf. chapitre 1). On les appellera une caractérisation des groupes $c_m(\mathbf{x}_i)$ pour le modèle m.

Dans la suite on présentera des techniques qui ne concernent pas exclusivement la discrimination mais aussi la régression ; ainsi $c_m(\mathbf{x}_i)$ sera plus généralement la sortie du modèle m et $y(\mathbf{x}_i)$ la variable à estimer par la combinaison de modèles. Elles sont évidemment de la même nature.

2.3 La combinaison par différenciation de modèles

La combinaison linéaire

Considérons une forme assez simple de combinaison des M modèles ; à savoir une combinaison linéaire

$$y_{COM}(\mathbf{x}) = \sum_{m=1}^{M} \beta_m c_m(\mathbf{x}). \quad (2.1)$$

Dans la mesure où les $c_m(\mathbf{x})$ sont connus, il ne reste qu'à trouver les poids β_m attribués à chaque modèle. Et c'est là qui se joue tout "l'art" de la combinaison. Plusieurs auteurs se sont interessés à la combinaison linéaire. On les découvrira au fil de ce chapitre.

L'idée de la combinaison linéaire a l'avantage d'être simple, naturelle et expressive. La variable expliquée se présente comme la somme du modèle combiné plus une erreur,

$$y(\mathbf{x}) = y_{COM}(\mathbf{x}) + \varepsilon(\mathbf{x}). \quad (2.2)$$

Naturellement on voudra minimiser l'expression E de la somme des erreurs quadratiques $E = \sum_{i=1}^{n} \varepsilon^2(\mathbf{x}_i)$. Alors il apparaît que, sous des hypothèses convenables, les meilleurs estimations des poids β_m sont obtenus par la méthode des moindres carrés.

La validation croisée

Deux problèmes se posent si l'on résout la combinaison par les moindres carrés. L'un est que l'on utilise les mêmes mesures $\mathbf{x}$ pour calculer la sortie de chaque modèle $c_m(\mathbf{x})$ et pour minimiser l'erreur E. Ce qui fait que les estimateurs de β_m sont biaisés. Pour surmonter cette difficulté, une approche consiste à ce que les $c_m(\mathbf{x})$ soient obtenues par validation croisée. On les notera $c_m(\mathbf{x})_{-i}$ pour souligner que l'observation i est absente au moment de l'estimation de $c_m(\mathbf{x})$. Ainsi (2.1) devient,

$$y_{COM}(\mathbf{x}_i) = \sum_{m=1}^{M} \beta_m c_m(\mathbf{x})_{-i}. \tag{2.3}$$

Il faut noter ici l'utilisation de la validation croisée avec une intention tout à fait différente de son utilisation habituelle. La validation n'est plus le but mais par contre on utilise la validation croisée pour construire un nouveau estimateur. Stone (1974) a appelé la restructuration (2.3) le "model-mix" et l'a illustré dans le cas d'une combinaison de modèles de régression où chaque équation de régression avait été construite avec un nombre différent de variables.

Dans une toute autre terminologie, propre aux réseaux de neurones, Wolpert (1992) propose le "stacked generalisation" pour combiner des modèles. Les notions sont pourtant exactement les mêmes de celles de "model-mix". Pour cet auteur, les $c_m(\mathbf{x})$ sont les estimations de niveau zéro et $c_m(\mathbf{x})_{-i}$ sont les estimations de niveau un. Son idée est que les $c_m(\mathbf{x})_{-i}$ contiennent plus d'information et peuvent être alors utilisés avantageusement pour construire un bon modèle combiné. Il l'a montré en formant des combinaisons non linéaires de modèles des voisins les plus proches. Mais il a remarqué que la façon d'utiliser les données de niveau un pour avoir des combinaisons correctes était un problème ouvert. Plus tard il donnera une méthode précise, le "stacking" (Smyth and Wolpert 1999) - sur laquelle on reviendra - pour déterminer les poids β_m optimaux dans le cas de la combinaison linéaire.

Les travaux de Wolpert ont été transposés en statistique par Breiman (1996b) qui apparemment méconnaissait le "model-mix". Souvent la dimension des données est grande en individus et en variables et alors la programmation de la validation croisée par unité est lourde et la méthode est très lente. D'après ses simulations Breiman (1996b) montre que la validation croisée par blocs est une méthode efficace. Bien que dans le texte la notation $-i$ souligne l'absence d'une seule observation, la validation croisée sera considerée aussi bien par unité que par blocs. La référence explicite à l'une d'entre elles sera faite si nécessaire.

La colinéarité

Le deuxième problème qui s'avère plus important car plus néfaste est la colinéarité. En effet il est très probable que les $\{c_m(\mathbf{x})\}$ sont colinéaires et correlées en conséquence du fait qu'elles modélisent les mêmes données. La colinéarité mesure la dépendance linéaire entre les modèles. Dans ce cas la méthode des moindres carrés produit des estimations de β_m peu stables puisque leur variance est grande. Ainsi les poids β_m obtenus sur un échantillon en particulier ne seront pas optimaux. Et ceci n'est pas résolu par la validation croisée, telle qu'on l'a considérée à la dernière section, en utilisant $\{c_m(\mathbf{x})_{-i}\}$. Pour résoudre ce problème plusieurs approches ont été développées.

L'un des moyens traditionnels pour traiter la colinéarité est d'imposer des contraintes. Déjà Stone (1974) pour des raisons intuitives avait indiqué l'intérêt d'étudier l'effet sur le "model-mix" de contraintes de non négativité ($\beta_m \geq 0$) et de somme unitaire ($\sum_m \beta_m = 1$).

Breiman (1996b) dans la même ligne propose d'effectuer une régression ridge.

De cette manière l'erreur E est minimisée assujetti à la contrainte $\sum_m \beta_m^2 = s$, s étant calculé par validation croisée. Il se rend compte, d'après un grand nombre de simulations, qu'il est préférable d'imposer $\beta_m \geq 0$. Cette méthode fondée sur le travail de Wolpert a été nommée "stacked regression".

Très intéressant est le théorème (Breiman 1996b - p.52) où l'auteur définit la condition sous laquelle la combinaison de modèles n'améliore pas l'erreur du meilleur modèle simple. Considérons l'estimation du modèle m,

$$y(\mathbf{x}) = c_m(\mathbf{x}) + \varepsilon_m(\mathbf{x}) \tag{2.4}$$

et ajoutons la contrainte de somme unitaire $\sum_m \beta_m = 1$. On définit les matrices R_{ml} des résidus croisés $R_{ml} = \sum_i \varepsilon_m(\mathbf{x}_i)\varepsilon_l(\mathbf{x}_i)$. L'expression de E peut être réécrite en fonction des m erreurs individuels $\varepsilon_m(\mathbf{x})$.

$$E = \sum_{m,l} \beta_m \beta_l R_{ml}, \quad m,l = 1,\ \ldots,\ M \tag{2.5}$$

L'auteur montre que le meilleur modèle $c_m(\mathbf{x})$ est aussi bon qu'une combinaison linéaire de modèles si et seulement si $R_{mm} \leq R_{lm}$, pour tous les l.

De plus il vérifie, numériquement, deux constatations. L'une, très importante, dit que plus de bénéfices adviennent d'une combinaison de modèles dissemblables. L'autre est que la contrainte $\sum_m \beta_m = 1$ n'est pas nécessaire car il semble que les β_m optimaux obtenus par minimisation non négative ont la propriété de $\sum_m \beta_m \simeq 1$. Toutefois l'expression (2.5) exige cette contrainte.

D'une façon innovatrice Breiman a appliqué le "stacking regression" à la combinaison d'arbres de différentes tailles qui résultent de l'algorithme CART (Breiman, Friedman, Olshen, and Stone 1984). Les $c_m(\mathbf{x})$ utilisent la portion de l'arbre avec m noeuds terminaux. Si $\mathbf{x}$ est M-dimensionnel, $c_m(\mathbf{x})$ est la sortie de la régression fondé sur les m variables qui ont la plus petite l'erreur. L'application de cette méthode à deux bases de données réelles a réduit l'erreur de l'ordre de 10%. Breiman a encore réalisé de nombreuses simulations sur la combinaison de régressions. Les modèles à combiner sont obtenus par sélection de variables au moyen d'une procédure de supression à chaque pas ou alors il combine des régressions ridge, chacune correspondant à un paramètre ridge différent. L'auteur a aussi combiné les deux types de régressions. Dans ses études jamais la combinaison n'est pire, au sens de l'erreur de validation croisée, que le meilleur modèle simple.

Un cas particulier de combinaison linéaire est la simple moyenne. Cela veut dire que les poids sont égaux pour tous les modèles, $\beta_m = \frac{1}{M}$. C'est certainement une méthode très naïve mais elle peut devenir intéressante en ayant un statut de référence par rapport à d'autres méthodes. Perrone and Cooper (1973) l'ont considérée et appelée "basic ensemble method". Soit la somme des erreurs quadratiques individuelles $E_m = \sum_i \varepsilon_m^2(\mathbf{x}_i)$. Les auteurs ont démontré qu'en moyennant les modèles, E est réduite d'un facteur M par rapport à $\frac{1}{M}\sum_m E_m$. En augmentant le nombre de modèles dans la combinaison l'erreur d'estimation serait arbitrairement petite. Pourtant les hypothèses propres à "basic ensemble method" (erreurs ε_m mutuellement indépendantes avec moyenne nulle) sont très souvent irréalistes ; son absence de réalisme limite beaucoup la portée de ce résultat.

Ces auteurs ont encore étudié le cas plus général de la combinaison linéaire qu'ils ont appellé "generalized ensemble method". Leur méthode détermine les poids optimaux en permettant la corrélation des erreurs ε_m et en permettant aussi que la moyenne de ε_m soit différente de zéro. On reprend (2.5) pour la minimiser sous la contrainte $\sum_m \beta_m = 1$. Par la technique classique des multiplicateurs de Lagrange les poids optimaux sont,

$$\beta_m = \frac{\sum_l R_{ml}^{-1}}{\sum_h \sum_l R_{hl}^{-1}}, \qquad m, l, h = 1, \ \ldots, \ M. \tag{2.6}$$

Si les ε_m ne sont pas correlées et si ils ont une espérance nulle (le cas de "basic ensemble method"), alors (2.6) est de la forme $\beta_m = \frac{R_{mm}^{-1}}{\sum_l R_{ml}^{-1}}$.

Cependant en présence de colinéarité il est impossible de calculer les inverses R_{ml}^{-1}. Pour ces "ensemble methods" Perrone et Cooper proposent une méthode de sélection de modèles pour alléger les effets de la colinéarité entre paires de modèles. Un modèle est sélectionné quand la corrélation avec un autre modèle dépasse un certain seuil. Le seuil proposé par les auteurs dépend manifestement du problème à résoudre.

Dans la même veine des derniers auteurs, Bishop (1995) reproduit la méthode "generalized ensemble method" en l'appelant comités de réseaux. Celeux, Brito, and Ferreira (1999) l'ont adaptée à la discrimination en faisant bien la distinction entre le cas des variables quantitatives et qualitatives. Hashem and Schmeiser (1995) ont développé une forme de combinaison comparable à "generalized ensemble method".

Afin de lutter contre la colinéarité Merz and Pazzani (1999) ont proposé l'algorithme PCR*. Cet algorithme répond aussi à la question importante d'attribuer une pondération unique à un certain modèle. À propos de ce problème ils affirment que : "Un réseau de neurones peut détecter des interactions non linéaires entre les variables originales tandis qu'une méthode de régression linéaire standard peut utiliser une stratégie de sélection pour simplifier la prévision. Une bonne combinaison doit attribuer une contribution unique à chaque modèle."

Si PCR* exploite tous les modèles disponibles, dans un autre article Merz (1999) réalise une Analyse de Correspondances sur un ensemble de modèles. Ensuite il cherche itérativement un sous-ensemble de modèles peu corrélés qui seront à la fin combinés par stacking.

Un moyen de s'affranchir de la colinéarité est de construire des modèles qui fassent des erreurs décorrelés en ajustant le biais du modèle (Opitz and Shavlik 1996) ou alors les données (Meir 1995). Pourtant ces deux astuces n'éliminent pas la colinéarité même s'ils la réduisent. Opitz et Shavlik accordent à chaque modèle une pondération d'après une estimation de sa précision a_m calculée sur un échantillon-test. Les poids sont déterminés par l'expression $\beta_m = \frac{1-a_m}{\sum_{u=1}^{M}(1-a_u)}$. Intuitivement on remarque que le modèle m a une plus forte contribution à mesure que sa performance estimée augmente par rapport à la performance estimée cumulée des autres modèles. Cette approche est, certes, moins sensible à la colinéarité mais moins performante car les corrélations entre modèles ne sont pas prises en compte.

Une autre tactique pour luter contre la colinéarité est d'utiliser des procédures du gradient descendant (Kivinen and Warmuth 1997). Celles-ci cherchent les coefficients avec des actualisations itératives des poids multiplicatifs ou exponentiels. Les actualisations sont fonction de sa performance sur l'échantillon test. Ces techniques évitent l'inversion de la matrice R. Pourtant ces procédures sont susceptibles d'être capturés par des minimum locaux et, de plus, il est difficile de choisir les poids initiaux ou de décider de combien il faut actualiser les poids.

Partant de l'estimateur classique des moindres carrés habituel

$$\tilde{\beta}_m = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T y \tag{2.7}$$

de β_m, où $\mathbf{C}$ est une matrice nxM, dont l'élément générique est $\{c_m(\mathbf{x})\}$, LeBlanc and Tibshirani (1996) ont proposé et analysé quelques méthodes pour estimer les β_m

- les quantités $(\mathbf{C^T C})$ et $\mathbf{C^T y}$ sont corrigés du biais en les divisant par n et en leur ajoutant d'une quantité calculée par bootstrap ;
- les $\{c_m(\mathbf{x})_{-i}\}$ sont approximés comme en validation croisée généralisée. Mais si les $\{c_m(\mathbf{x})\}$ viennent d'une estimation adaptivement choisie cette manière de faire n'est pas valide.

Néanmoins, ces deux méthodes ne donnent pas de bons résultats d'après les simulations des auteurs. Ainsi ils ont décidé d'ajouter une régularisation des estimations.

- Les β_m sont rapprochés par rétrécisement vers le vecteur $\{\frac{1}{M}\}^M$. Ainsi $\tilde{\beta}_m = (\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I})^{-1}(\mathbf{C}^T y + (\frac{\lambda}{M})\mathbf{1})$, λ étant un facteur fixe de 75%.

Les auteurs ont encore proposé une autre méthode de combinaison, d'esprit différent des précédentes, mais assez séduisante. Les $\tilde{\beta}_m$ seraient considérés comme une mesure de performance du modèle m. Par exemple

$$\tilde{\beta}_m = \frac{L(\hat{\theta}_m, c_m(\mathbf{x}))}{\sum_{u=1}^{M} L(\hat{\theta}_u, c_u(\mathbf{x}))} \tag{2.8}$$

où $L(\hat{\theta}_m, c_m(\mathbf{x}))$ est la vraisemblance pour le modèle m.

Dans le même article, LeBlanc et Tibshirani ont considéré le problème de la combinaison en l'analyse discriminante. Dans la pratique il est préférable de réécrire z_i comme un vecteur multivarié de dimension K. Ainsi pour chaque observation i on dispose d'un vecteur où sa composante k vaut 1 si l'observation appartient au groupe k et 0 sinon. La combinaison se fait avec les probabilités estimées des groupes $p_m(\mathbf{x})$. D'après leurs exemples la combinaison d'une méthode de discrimination linéaire et d'une méthode de voisins les plus proches prouve son efficacité si elle est accompagnée de la contrainte de non négativité des poids.

Le point de vue bayésien

La combinaison de modèles peut être vue sous une formulation bayésienne. La moyenne prédictive d'une nouvelle observation z_0 caractérisée par le vecteur prédictif $\mathbf{x}_0$ s'exprime

$$E(z_0 \mid \mathbf{t}_0, \mathbf{x}_0) = \int E(z_0 \mid x_0, \theta, m, \beta) p(\theta, \beta, m \mid \mathbf{t}_0) d\theta d\beta dm \tag{2.9}$$

où m est le m-ème modèle et $\theta = (\theta_1, \ \ldots, \ \theta_M)$ sont les paramètres pour les modèles de la combinaison. La moyenne prédictive (2.9) peut être réécrite

$$E(z_0 \mid \mathbf{t}_0, \mathbf{x}_0) = \int \int [\sum_m \bar{c}(\mathbf{x}_0, \theta, m) p(m \mid \theta, \beta, \mathbf{t}_0)] \times p(\theta, \beta \mid \mathbf{t}_0) d\theta d\beta. \tag{2.10}$$

Cette formulation entraîne le respect des contraintes de non négativité des β_m ainsi que de leur somme unitaire.

Les coefficients variables

Et si les pondérations n'étaient pas constantes ? Deux autres schémas de combinaison permettent que les β_m varient avec $\mathbf{x}$,

$$\sum_m \beta_m(\mathbf{x}) c_m(\mathbf{x}) \tag{2.11}$$

où alors les β_m varient seulement avec les $c_m(\mathbf{x})$ soit

$$\sum_m \beta_m(c_m(\mathbf{x})c_m(\mathbf{x})) \tag{2.12}$$

Ces deux modèles sont du type de coefficients variables (Hastie and Tibshirani 1993). LeBlanc and Tibshirani (1996) suggèrent que ces modèles seraient bien dans le cas des 'scatterplot smoothing'. Mais ce type d'approche est peu parcimonieux.

2.4 La combinaison de modèles par l'approche modèle unique

Un autre mode de combinaison de modèles sont les stratégies de rééchantillonnage. Chaque modèle est reproduit avec le même algorithme mais sur différents échantillons d'apprentissage. L'objectif est de tirer parti d'un ensemble de méthodes qui vraisemblablement feront des erreurs non corrélés. Les données pour un modèle en particulier sont obtenues par tirage d'un échantillon des données initiales selon une densité. Le *bagging* (Breiman 1996a) et le *boosting* (Schapire 1990) sont deux techniques qui diffèrent par la densité de tirage utilisée.

Le bagging est une méthode pour exploiter la variance d'un modèle en l'appliquant à plusieurs versions de l'ensemble de données. Les variations sur l'échantillon sont obtenues avec répétition. La probabilité qu'un exemple soit retiré est uniforme et le nombre d'exemples retirés est égale à la taille de l'échantillon original. Cette approche suppose que les modèles doivent être pondérés uniformément. Breiman (1996a) conclut que bagging donne de bons résultats pour des procédures instables. Mais il peut légèrement dégrader la performance de procédures stables. Les procédures instables sont les réseaux de neurones, les arbres de classification et de régression, la sélection de variables en régression linéaire tandis que les méthodes des voisins les plus proches ou la discrimination linéaire sont des procédures stables.

Une autre technique de rééchantillonnage est fondée sur le boosting. Cette technique estime qu'un modèle peu précis peut être estimé plusieurs fois de suite pour aboutir à une règle de prévision assez précise. Freund and Schapire (1996) ont développé et raffiné cette approche en la mettant en œuvre. Cette technique attribue un poids à chaque élément de l'échantillon et l'ajuste après l'estimation du modèle. Au début les points sont pondérés uniformément. Ensuite les points ayant une petite erreur de prévision prennent un petit poids et les points ayant une grande erreur de prévision prennent un grand poids. L'idée est d'obliger le modèle à se concentrer sur les points de l'échantillon plus difficiles à classer. Les données sont rééchantillonnés avec répétition selon la distribution des poids des points. Les auteurs ont décrit un algorithme de combinaison, Adaboost, où la pondération du modèle à l'étape t, $t = 1, \ldots, T$, (T est le nombre de fois que le modèle est réestimé) est fonction de l'erreur de classement à l'étape t : $\beta_t = \log \frac{(1-\epsilon_t)}{\epsilon_t}$, avec ϵ_t représentant le taux d'erreur à cette étape.

2.5 La combinaison linéaire dynamique

Une approche fonctionnelle d'estimation des β_m a été proposée par Jacobs, Jordan, Nowlan, and Hinton (1991) et par Tresp and Taniguchi (1995). Dans ce cadre chaque modèle a un poids qui est variable selon la partie de l'échantillon en jeu.

La plus connue et intéressante des approches dynamiques est la "mixture of experts" de Jacobs, Jordan, Nowlan, and Hinton (1991). Plusieurs modèles experts plus un réseau huissier qui décide lesquels des experts devront être utilisés pour chaque cas. Chaque expert donne une densité cible pour un exemple donné. Le réseau huissier sélectionne un ou plusieurs experts qui semblent avoir la densité cible la plus appropriée. Pendant la procédure d'estimation les changements de poids se font sur les experts choisis. Les experts qui sont plus précis pour cet exemple se voient attribuer plus de responsabilités et les autres moins. Les poids des autres experts, qui se spécialisent dans beaucoup de cas différents, ne sont pas modifiés. Les experts deviennent focalisés parce que ses poids sont découplés des poids des autres experts et il finiront par se spécialiser dans une petite portion de l'espace d'entrée.

Jordan and Jacobs (1994) ont étendu cette approche en considérant les modèles experts pour les modèles linéaires généralisés. Les experts sont des feuilles dans une architecture en arbre dont les nœuds internes sont des fonctions huissiers. Ces fonctions font des coupures légères permettant que les données appartiennent simultanément à plusieurs régions.

Tresp and Taniguchi (1995) ont présenté une collection de fonctions pondératrices qui peuvent être utilisées pour combiner les modèles. Ces méthodes pondérent un modèle d'après sa position sur la région de confiance de l'exemple ou la quantité de variabilité des prévisions du modèle dans cette région. Ces approches exigent la non négativité des poids ainsi que leur somme soit égale à un. Mais ils n'ont pas été encore évalués empiriquement.

Chapitre 3

Notre approche de la combinaison de modèles

On applaudit
avec les deux mains.

Dans ce chapitre on analyse en détail les méthodes de combinaison qui nous concernent. Comme on l'a écrit précédemment il y a essentiellement deux approches de combinaison, l'approche modèle unique et l'approche differenciation de modèles. Rappelons que selon la première chaque modèle est reproduit avec le même algorithme. Elle s'appuie sur des stratégies de reéchantillonnage tels que le *bagging* ou le *boosting*. Pour la deuxième, les modèles composants de la combinaison sont tous fondés sur le même échantillon mais sont formellement différents ou ils sont du même genre mais diffèrent en complexité. C'est cette approche qui nous intéresse en particulier et que nous développons dans la suite.

Dans une combinaison de modèles de type differenciation de modèles on peut distinguer deux types : la combinaison en série, en général linéaire, et la combinaison hiérarchique. Si la combinaison est du type combinaison en série les modèles sont considérés au même niveau selon une certaine forme dont la plus courante est la combinaison linéaire. Au long de ce chapitre on présente plusieurs méthodes de combinaison linéaire qui reposent soit sur le critère des moindres carrés (section 3.1.1) soit sur la vraisemblance des modèles (section 3.2). Dans une combinaison du type hiérarchique les modèles sont considérés à différents niveaux emboîtées. La combinaison hiérarchique est analysée dans la section 3.3.

En analyse discriminante la population est constituée de K groupes connus et est caracterisée par d variables $\{X_1,\ X_2,\ \ldots,\ X_d\}$. L'échantillon d'apprentissage est l'ensemble des n observations indépendantes $\mathbf{t}_i = (\mathbf{x}_i, z_i)$, $i = 1,\ \ldots,\ n$, où $\mathbf{x}_i$ sont les réalisations des d variables pour l'observation i et z_i son groupe d'appartenance. Par hypothèse z_i est la réalisation d'une autre variable aléatoire Z. Le modèle d'analyse discriminante est caractérisé par les densités conditionnelles au groupe d'appartenance $f_k(\mathbf{x})$, les probabilités a priori des groupes π_k et les probabilités a posteriori $p_k(\mathbf{x})$.

Nous plaçons la combinaison de modèles dans le contexte gaussien des modèles EDDA. Un incursion non paramétrique est faite en tenant compte dans la combinaison de la méthode des k plus proches voisins (k-ppv). Les modèles considerés conduisent naturellement à l'estimation des densités par groupe $f_k(\mathbf{x})$. Les côuts de mauvais classement pour les différents groupes sont supposés identiques et unitaires. Les probabilités a priori des groupes sont estimés par $\frac{n_k}{n}$ où n_k est le nombre d'individus qui appartient au groupe k. La règle d'affectation aux groupes est définie par la règle de Bayes habituelle

$$\arg\max_k \pi_k p_k(\mathbf{x}),\ \ k = 1,\ \ldots,\ K.$$

Il se pose tout d'abord la question de savoir si la combinaison doit se fonder sur les décisions des modèles, c'est-à-dire le groupe prédit, ou sur les densités estimées par les modèles. Le produit final d'une analyse discriminante est l'affectation à un groupe, aussi la combinaison pourrait se faire sur les décisions déduites des modèles. Mais pour les méthodes de combinaison que nous développons il est plus pertinent d'utiliser les densités car elles sont plus riches en information.

Les modèles dans la combinaison sont alors caractérisés par les densités par groupe. Plusieurs stratégies de caractérisation sont envisagées. On peut construire la combinaison à partir

- des densités conditionnelles $f_k(\mathbf{x})$,
- du ratio des densités conditionnelles par rapport à un groupe de référence $\frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}$ (l est le groupe de référence),
- du logarithme de ce ratio $\log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}$
- ou encore des probabilités a posteriori $p_k(\mathbf{x})$.

Nous verrons que, d'un point de vue théorique, les méthodes de combinaison présentées ne travaillent pas de la même façon à partir de toutes ces caractérisations.

3.1 La combinaison en série

Considérons une combinaison linéaire de M modèles d'analyse discriminante. Schématiquement elle a la forme

$$p_{COM}(\mathbf{x}) = \sum_{m=1}^{M} p_m(\mathbf{x})\beta_m \tag{3.1}$$

où $p_{COM}(\mathbf{x})$ représente les probabilités à estimer par combinaison, $p_m(\mathbf{x})$ sont les probabilités obtenues par application du modèle m et β_m le poids attribué au modèle m dans la combinaison.

Dans l'expression (3.1) la référence au groupe est délibérément omise. La combinaison de modèles en série a surtout été utilisée en régression (Perrone and Cooper 1973, Stone 1974, Bishop 1995, Hashem and Schmeiser 1995, Breiman 1996b) où la sortie de chaque modèle est, pour chaque observation, unidimensionnelle. Dans le cas des modèles d'analyse discriminante les sorties sont, par nature, K-dimensionnelles. Les méthodes de combinaison doivent s'adapter à cette particularité qui joue un rôle important dans les techniques de la section 3.1.1.

Les $p_m(\mathbf{x})$ sont estimées préalablement à l'analyse par combinaison. Dans le présent travail ces probabilités sont les sorties des modèles EDDA et de la méthode des k plus proches voisins. Ainsi les $p_m(\mathbf{x})$ sont connues et pour calculer les $p_{COM}(\mathbf{x})$ il ne reste qu'à obtenir les bons coefficients β_m. Par hypothèse ces pondérations sont indépendantes des observations $\mathbf{x}$.

On présente ici deux types de méthodes de combinaison classés selon leur fondement théorique. Il y a les méthodes qui visent à minimiser l'erreur d'ajustement aux moindres carrées (sections 3.1.2, 3.1.3 et 3.1.4) et les méthodes qui reposent sur la vraisemblance (sections 3.2.1, 3.2.2, 3.2.3 et 3.2.4).

3.1.1 Les moindres carrés

Considérons les vraies probabilités inconues $p(\mathbf{x})$ et que l'on veut approximer par $p_{COM}(\mathbf{x})$. On peut exprimer leur rapport comme,

$$p_{COM}(\mathbf{x}) = p(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{3.2}$$

où $\varepsilon(\mathbf{x})$ représente l'erreur d'ajustement des données au modèle. Pour trouver de "bons" coefficients de combinaison β_m dans le cadre décrit par (3.1), la méthode la plus naturelle est la méthode des moindres carrés qui mimimise la somme du carré des erreurs d'ajustement $\sum_{i=1}^{n} \varepsilon^2(\mathbf{x}_i)$.

La variable de réponse est le vecteur y_k, $(k = 1, \ldots, K)$ qui pour chaque observation vaut 1 si l'observation appartient au groupe k et 0 sinon. La matrice des régresseurs $\mathbf{P}(\mathbf{x})$ est construite avec les probabilités a posteriori des K groupes estimées par les M modèles, $p_m^k(\mathbf{x})$.

La nature K-dimensionnelle des sorties des modèles d'analyse discriminante influence la façon de construire la matrice $\mathbf{P}(\mathbf{x})$. C'est ainsi que trois différentes structures de $\mathbf{P}(\mathbf{x})$ sont associées aux trois stratégies de combinaison par les moindres carrés que nous présentons dans cette section.

Parmi les caractérisations considérées[1] nous avons choisi les probabilités a posteriori $p_m^k(\mathbf{x})$ pour caractériser les modèles. Cela s'explique par le fait que

[1] Les densités conditionnelles $f_k(\mathbf{x})$, le ratio des densités conditionnelles par rapport à un groupe de référence $\frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}$, le logarithme de ce dernier ratio $\log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}$ ou encore les probabilités a posteriori $p_k(\mathbf{x})$.

les vecteurs réponse y_k sont des vecteurs indicateurs binaires. En effet, les $p_m^k(\mathbf{x})$ prennent leurs valeurs dans $[0,1]$ alors que les $f_m^k(\mathbf{x})$ pour les modèles considérés ici ne sont pas limités à $[0,1]$, car dans la présente thèse les modèles de référence sont les modèles gaussiens EDDA.

L'application des moindres carrés dans le contexte de (3.1) ne se fait pas sans difficultés. En effet, deux problèmes surgissent : l'un concerne le biais des estimateurs des β_m et l'autre regarde la colinéarité des régresseurs $\mathbf{P}(\mathbf{x})$.

L'échantillon d'aprentissage $\mathbf{t}$ est communément utilisé pour estimer $p_m(\mathbf{x})$ et ensuite estimer les coefficients β_m. Une telle procédure conduit à des estimateurs biaisés des β_m car la nécessaire indépendance des $p_m(\mathbf{x})$ n'est pas assurée.

Un moyen d'éliminer le biais est d'utiliser des échantillons différents pour les deux tâches (estimation de $p_m(\mathbf{x})$ et de β_m), ceux-ci pouvant être constitués par validation croisée. Ainsi, nous remplaçons $p_m^k(\mathbf{x})$ par $p_m^k(\mathbf{x})_{-i}$ où l'indice $_{-i}$ signifie que la densité de $\mathbf{x}_i$ a été estimée en enlevant i de l'échantillon d'aprentissage,

$$p_m^k(\mathbf{x})_{-i} \propto f_m^k(\mathbf{x})_{-i} = f^k(\mathbf{x}_i|\hat{\theta}_{-i}, m), \quad i = 1, ..., n. \tag{3.3}$$

La validation croisée est ici employée dans une acceptation différente de son utilisation habituelle en analyse discriminante. Nous ne sommes pas concernés par la validation mais par l'élaboration d'échantillons d'apprentissage différents. Sans expliciter cette problématique, Wolpert (1992) atteste la nécessité de l'utilisation de ce qu'il appelle les données de "niveau un", par opposition à l'échantillon original dénommé les données de "niveau zéro".

Une deuxième difficulté qui se pose pour l'utilisation des moindres carrés dans les circonstances de (3.1) est la colinéarité. Un nombre p de variables sont colinéaires ou dépendantes linéairement si l'un des vecteurs représentatifs des variables peut se traduire par une combinaison linéaire des autres.

La colinéarité entraîne de graves conséquences pour l'utilisation des moindres carrés. Rappellons que la solution des moindres carrés pour les estimateurs des β_m est

$$\tilde{\beta} = (\mathbf{P}(\mathbf{x})^T \mathbf{P}(\mathbf{x}))^{-1} \mathbf{P}(\mathbf{x})^T y_k, \tag{3.4}$$

où $\tilde{\beta} = (\tilde{\beta}_1, \ \ldots, \ \tilde{\beta}_M)$. Or en présence de colinéarité la matrice $\mathbf{P}(\mathbf{x})^T \mathbf{P}(\mathbf{x})$ n'est plus inversible. Dans des cas moins extrêmes la matrice est inversible mais très proche de la singularité (le déterminant de $\mathbf{P}(\mathbf{x})^T \mathbf{P}(\mathbf{x})$ avoisine zéro). En conséquence, la solution manque de stabilité.

La colinéarité des régresseurs cause une forte corrélation des coefficients de régression. Aussi on ne peut plus interpréter les coefficients de régression comme une mesure de la sensibilité de la variable réponse à des variations des régresseurs. S'il y a colinéarité les variances estimées des estimateurs des β_m est grande. Dans ces conditions les coefficients de régression ont une grande variabilité d'un échantillon à l'autre et peuvent individuellement ne pas être statistiquement significatifs malgré la liaison physique entre les régresseurs et la variable réponse.

Pour s'affranchir de la colinéarité, il existe plusieurs stratégies dont voici les principales et les moins controversés (Pour un bon panorama sur cette question voir Hashem and Schmeiser 1995.),

- Introduire de nouvelles données de façon à annuler la structure de colinéarité.
- Utiliser la régression *ridge* (Hoerl 1970) comme procédure d'estimation.

– Abandonner quelqu'uns des régresseurs de façon à affaiblir la colinéarité.

Nous avons opté pour la régression *ridge* qui consiste à introduire des contraintes sur les coefficients de la régression. Il s'agit d'une voie intéressante d'attaque de la colinearité car elle garde toute l'information de l'échantillon. La régression *ridge* est basée sur le critère des moindres carrés sous la contrainte $\sum_m \beta_m^2 = s$. La valeur optimale de s peut être sélectionnée par validation croisée. Breiman (1996b) a examiné la performance de cette méthode[2] mais les résultats s'avèrent ne pas être consistants. Ainsi, nous utilisons des contraintes de non négativité des coefficients $\beta_m \geq 0$ et de somme unitaire $\sum_m \beta_m = 1$. On les désignera dans la suite, respectivement, contrainte de non négativité et contrainte unitaire.

L'introduction de ces contraintes est un moyen d'atténuer les effets de la colinéarité. D'ailleurs ces contraintes sont naturels : il est normal que chaque modèle soit donne une contribution positive dans la combinaison en série soit donne une contribution nulle.

La solution des moindres carrés pour les estimateurs des β_m sous la contrainte de non négativité est obtenue par un algorithme de Lawson and Hanson (1974) programmé en Matlab©. Sous la contrainte unitaire la solution des moindres carrés pour les estimateurs des β_m est obtenue par recours aux multiplicateurs de Lagrange. La solution est

$$\tilde{\beta} = \mathbf{AB} + \mathbf{A1}^T(\mathbf{1A1}^T)^{-1}(1 - \mathbf{1AB}) \tag{3.5}$$

(Judge, Griffiths, Hill, and Lee 1980) où $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_M)$, $\mathbf{A} = (\mathbf{P}(\mathbf{x})^T\mathbf{P}(\mathbf{x}))^{-1}$, $\mathbf{B} = \mathbf{P}(\mathbf{x})^T y_k$ et $\mathbf{1}$ est un vecteur unitaire de dimension M.

[2]Dans un contexte où les modèles à combiner sont des modèles de régression.

Nous présentons maintenaint trois méthodes classiques de combinaison qui reposent sur le critère des moindres carrés. Ces techniques se distinguent par l'expression du critère d'erreur à minimiser. Nous envisageons les problématiques suivantes,

- une régression séparée pour chaque groupe ;
- la vectorisation de la matrice des probabilités ;
- le comité de méthodes.

3.1.2 Une régression séparée pour chaque groupe

La formule de combinaison de modèles en série a, dans ce cas spécifique, l'expression

$$p_{COM}^{k}(\mathbf{x}) = \sum_{m} p_{m}^{k}(\mathbf{x})\beta_{m}^{k}, \quad k = 1, \ \ldots, \ K. \tag{3.6}$$

Les probabilités $p_m^k(\mathbf{x})$ sont les probabilités obtenus par validation croisée mais pour simplifier les notations nous omettons ici et dans la suite l'indice $-i$.

La matrice des régresseurs s'écrit

$$\mathbf{P}(\mathbf{x}) = \left[p_1^k(\mathbf{x}) \ p_2^k(\mathbf{x}) \ \ldots \ p_M^k(\mathbf{x})\right].$$

Dans ce contexte l'expression de l'erreur de régression est

$$e_{i,k} = y_{i,k} - \sum_{m=1}^{M} p_{m}^{k}(\mathbf{x}_i)\beta_{m}^{k}, \quad k = 1, \ldots, K. \tag{3.7}$$

L'application du critère des moindres carrés conduit à minimiser $\sum_{i=1}^{n} e_{i,k}^2$. Des contraintes de non négativité sont introduites. L'erreur de régression $e_{i,k}$ est l'expression particulière à cette méthode de combinaison de $\varepsilon(\mathbf{x})$ en (3.2).

Pour les problèmes à plus que deux groupes, on a intérêt, pour des raisons d'interprétation, à faire une construction dichotomique et à considérer un groupe contre tous les autres. C'est cette approche que nous adoptons ici. Remarquons que la variable réponse est un vecteur indicateur binaire de sorte que les régresseurs prennent en compte seulement le comportement du groupe k.

Par cette méthode K régressions sont effectuées et on en tire MxK coefficients de combinaison. Par conséquent pour un même modèle les groupes sont pondérés différemment dans la combinaison. En fait, ce nombre excessif de poids des modèles est un désavantage de cette technique, notamment en ce qui concerne l'interprétation de la combinaison.

3.1.3 Vectorisation de la matrice des probabilités

La matrice $\mathbf{P}(\mathbf{x})$ est construite avec les $p_m^k(\mathbf{x})$ juxtaposées de la façon suivante,

$$\begin{aligned}\mathbf{P}(\mathbf{x}) = \ & [p_1^1(\mathbf{x})\ p_1^2(\mathbf{x})\ \ldots\ p_1^K(\mathbf{x})\ p_2^1(\mathbf{x})\ p_2^2(\mathbf{x})\ \ldots\ p_2^K(\mathbf{x}) \\ & \ldots p_M^1(\mathbf{x})\ p_M^2(\mathbf{x})\ \ldots\ p_M^K(\mathbf{x})\big].\end{aligned}$$

La combinaison souhaitée des M modèles s'écrit,

$$p_{COM}^j(\mathbf{x}) = \sum_m \sum_k p_m^k(\mathbf{x})\beta_m^k, \quad j = 1, \ldots, K. \tag{3.8}$$

L'erreur de régression est alors

$$e_{i,j}^* = y_{i,j} - \sum_{m=1}^{M} \sum_{k=1}^{K} p_m^k(\mathbf{x}_i)\beta_m^k, \ j = 1, \ldots, K \tag{3.9}$$

Les β_m^k sont obtenus par minimisation de $\sum_{i=1}^n e_{i,j}^{*2}$ avec la contrainte de non négativité des poids.

De même que dans l'analyse précédente le fait d'avoir K coefficients par modèle signifie que pour un même modèle les groupes sont pondérés diffèrement. Cela s'avère négatif à cause de la quantité de poids (MxK) à prendre en considération.

Remarquons que si $K > 2$, chaque régression représente la position d'un groupe par rapport à tous les autres. Au contraire de la méthode de régression séparée pour chaque groupe (section 3.1.2), dans cette analyse l'interaction entre les groupes est prise en compte.

3.1.4 Le comité de méthodes

Nous considérons ici une technique de combinaison introduite par Bishop (1995). Il l'a proposée pour combiner des réseaux de neurones dans un contexte de régression et il l'a appelée comités de réseaux. Mais il est possible de l'adapter au cadre de la discrimination. C'est ce que nous avons fait dans un premier texte, Celeux, Brito, and Ferreira (1999) et continuons ici, désignant cette méthode "comité de méthodes".

La technique du comité de méthodes utilise, comme les deux techniques précédentes, le critère des moindres carrés. La principale différence est l'expression d'erreur considérée. Pour les méthodes de régression séparée par groupe et de vectorisation de la matrice des probabilités l'approximation $\mathbf{p}_{COM}(\mathbf{x})$ aux vraies probabilités $\mathbf{p}(\mathbf{x})$ est donnée par la somme de $\mathbf{p}(\mathbf{x})$ avec l'erreur $\varepsilon(\mathbf{x})$. Le comité de méthodes prend en compte l'erreur d'approximation de chaque modèle m,

$$\mathbf{p}_m(\mathbf{x}) = \mathbf{p}(\mathbf{x}) + \varepsilon_m(\mathbf{x}). \tag{3.10}$$

En conséquence, le modèle de combinaison selon le comité de méthodes s'écrit,

$$\begin{aligned} \mathbf{p}_{COM}(\mathbf{x}) &= \textstyle\sum_m \mathbf{p}_m(\mathbf{x})\beta_m \\ &= \mathbf{p}(\mathbf{x}) + \textstyle\sum_m \varepsilon_m(\mathbf{x})\beta_m \end{aligned} \tag{3.11}$$

où les probabilités estimées par combinaison $\mathbf{p}_{COM}(\mathbf{x})$ ont la forme

$$\mathbf{p}_{COM}(\mathbf{x}) = \left[p^1_{COM}(\mathbf{x})\ p^2_{COM}(\mathbf{x})\ \ldots\ p^K_{COM}(\mathbf{x})\right],$$

et où les probabilités a posteriori des modèles s'expriment par

$$\mathbf{p}_m(\mathbf{x}) = \left[p^1_m(\mathbf{x})\ \ p^2_m(\mathbf{x})\ \ \ldots p^K_m(\mathbf{x})\right].$$

On introduit la matrice de corrélation des erreurs $\mathbf{C}$ dont l'élément générique est

$$C_{uv} = E[\varepsilon_u(\mathbf{x})\varepsilon_v(\mathbf{x})^T], \qquad u, v = 1, \ldots, M \tag{3.12}$$

où $E[.]$ signifie l'espérance mathématique.

Ainsi, pour la tecnhique du comité de méthodes, l'erreur que l'on voudra minimiser au sens des moindres carrées est

$$\sum_u \sum_v \beta_u \beta_v C_{uv}$$

avec la contrainte de somme unitaire. Par recours aux multiplicateurs de Lagrange la solution pour les estimateurs des β_m est

$$\tilde{\beta}_m = \frac{\sum_{v=1}^{M} (\mathbf{C}^{-1})_{mv}}{\sum_{u=1}^{M}\sum_{v=1}^{M} (\mathbf{C}^{-1})_{uv}}, \tag{3.13}$$

où $(\mathbf{C}^{-1})_{uv}$ est l'élément générique de la matrice inverse de $\mathbf{C}$.

La matrice $\mathbf{C}$ est estimée par

$$\tilde{C}_{uv} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{p}_u(\mathbf{x}_i))(\mathbf{y}_i - \mathbf{p}_v(\mathbf{x}_i))^T. \tag{3.14}$$

Pour faire le lien avec les techniques précédentes utilisant aussi le critère des moindres carrés, nous dirons que le comité de méthodes considère l'écart entre les densités par groupe des observations et les vecteurs binaires d'origine associés pour chaque modèle m

$$e_{i,k,m} = y_{i,k} - p_m^k(\mathbf{x}_i). \tag{3.15}$$

Partant de ces écarts, la corrélation entre les erreurs des modèles u et v est

$$\tilde{C}_{u,v} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} e_{i,k,u} e_{i,k,v}. \tag{3.16}$$

Les coefficients β_m sont obtenus par minimisation de $\sum_{u=1}^{M} \sum_{v=1}^{M} \beta_u \beta_v \tilde{C}_{u,v}$ sous la contrainte de somme unitaire.

Nous proposons deux approches de comités de méthodes. Si le comité est fondé sur les probabiités a posteriori alors l'approche est désignée "*fuzzy classification rule*". C'est la solution que nous préconisons. L'autre solution est de fonder le comité sur les décisions des modèles ce que conduit à l'approche "*hard classification rule*". Nous n'avons pas poursuivi cette approche qui, comme nous l'avons déjà exprimé, dans le cas des modèles gaussiens, conduit à des combinaisons peu intéressantes car fondées sur une information très restrictive des modèles en compétition. Par contre, c'est une approche utilisé par Ferreira (2001), dans le cas des variables qualitatives.

L'un des intérêts de la combinaison de modèles par comités est de proposer un coefficient indépendant des groupes a priori, ce qui facilite son interprétation. Cette particularité fait la distinction essentielle entre les comités de méthodes et la combinaison par régression groupe par groupe ou par la vectorisation des probabilités.

3.2 L'utilisation de la vraisemblance

Les techniques présentées dans cette section sont d'esprit différent des techniques précédentes car elles s'appuient sur la vraisemblance des modèles à combiner. Notons que ces techniques visent à proposer un seul coefficient de pondération par modèle, indépendamment des groupes à discriminer.

Dans ces conditions la combinaison des modèles est de la forme

$$\mathbf{p}_{COM}(\mathbf{x}) = \sum_m \mathbf{p}_m(\mathbf{x})\beta_m \tag{3.17}$$

où les probabilités estimées par combinaison $\mathbf{p}_{COM}(\mathbf{x})$ s'écrivent

$$\mathbf{p}_{COM}(\mathbf{x}) = \left[p^1_{COM}(\mathbf{x})\ p^2_{COM}(\mathbf{x})\ \ldots\ p^K_{COM}(\mathbf{x})\right]$$

et où les probabilités a posteriori des modèles s'expriment

$$\mathbf{p}_m(\mathbf{x}) = \left[p^1_m(\mathbf{x})\ \ p^2_m(\mathbf{x})\ \ \ldots p^K_m(\mathbf{x})\right].$$

Les coefficients β_m sont estimés à partir de la vraisemblance associée au modèle m. Considérons l'échantillon d'apprentissage $\mathbf{t}_i = (\mathbf{x}_i, z_i),\ i = 1,\ \ldots,\ n$ qui a été obtenu selon le schéma d'échantillonnage de mélange.

La fonction de vraisemblance $L(\theta, \Pi)$ avec $\Pi = (\pi_1, \ \ldots, \ \pi_K)$ s'écrit[3]

$$L_m(\theta, \Pi) = \prod_i \prod_k \left[f_m^k(\mathbf{x})\pi_k\right]^{y_{ik}}. \tag{3.18}$$

Pour chaque modèle m les paramètres θ sont estimés selon la reparametrisation EDDA. La vraisemblance du modèle est alors $L(\tilde{\theta}, \tilde{\Pi})$ évaluée pour la valeur estimée du paramétre θ, et π_k est estimé par $\frac{n_k}{n}$.

Remarque : Le modèle des k-ppv étant une méthode non paramétrique, la fonction de vraisemblance n'est pas définie et cette approche ne peut être utilisée lorsque ce type de méthodes fait partie des méthodes à combiner.

Dans les sections suivantes nous estimons les β_m respectivement par le ratio des vraisemblances, le ratio des vraisemblances pénalisées et le ratio des vraisemblances intégrées.

3.2.1 Les ratios de vraisemblance

LeBlanc and Tibshirani (1996) ont proposé une métode de combinaison assez simple, néanmoins ils ne l'ont jamais utilisée. Dans leur optique les coefficients de pondération des modèles dans la combinaison β_m sont estimés proportionnellement aux ratios de vraisemblance,

[3] Si l'échantillon est obtenu selon un schéma d'échantillonnage rétrospectif la vraisemblance est $L_m(\theta) = \prod_i \prod_k \left[f_m^k(\mathbf{x})\right]^{y_{ik}}$.

$$\tilde{\beta}_m = \frac{L_m(\theta; \mathbf{x})}{\sum_{u=1}^{M} L_u(\theta; \mathbf{x})} \tag{3.19}$$

où $L_m(\theta; \mathbf{x})$ représente la vraisemblance associée au modèle m. Pour éviter que $\mathbf{t}$ soit employé deux fois nous calculons la vraisemblance par validation croisée comme on l'a vu précédemment.

En utilisant ce ratio nous incorporons, implicitement, dans l'analyse, la contrainte de non négativité et la contrainte de somme unitaire des poids. Cette particularité donne plus de force à la technique de combinaison car elle allie sa simplicité à l'existence de propriétés de stabilité et de cohérence. Le fait d'avoir un seul poids par modèle signifie que tous les groupes pour chaque modèle sont pareillement pondérés.

L'inconvénient de cette méthode est qu'elle favorise les modèles les plus complexes, du fait que plus un modèle est complexe plus grande est sa vraisemblance. Aussi, les solutions proposées dans les deux sections suivantes sont préférables.

3.2.2 Les ratios de vraisemblance pénalisée

Afin de corriger la dépendance de la technique précédente vis-à-vis de la complexité des modèles nous avons introduit une pénalisation. Cette pénalisation utilise le nombre de paramètres de chaque modèle qui est une mesure de la complexité. Pour pondérer les vraisemblances avec le nombre de paramètres nous sommes partis du critère d'information de Akaike (AIC) (Akaike 1974)

$$AIC = -2 \log L_m(\theta; \mathbf{x}) + 2\nu_m \tag{3.20}$$

où ν_m représente le nombre de paramètres indépendants du modèle m et $L_m(\theta;\mathbf{x})$ sa vraisemblance. De (3.20), on déduit

$$e^{-\frac{AIC}{2}} = L_m(\theta;\mathbf{x}).e^{-\nu_m}.$$

Les coefficients β_m sont alors estimés par les ratios de vraisemblance pondérée où $L_m(\theta;\mathbf{x})$ est remplacée par $L_m(\theta;\mathbf{x}).e^{-\nu_m}$

$$\tilde{\beta}_m = \frac{L_m(\theta;\mathbf{x}).e^{-\nu_m}}{\sum_{u=1}^{M} L_u(\theta;\mathbf{x}).e^{-\nu_u}}. \tag{3.21}$$

Nous n'utiliserons pas la vraisemblance de validation croisée tel qu'on l'a fait pour les ratios de vraisemblance. En effet travailler sur la vraisemblance par validation croisée n'est pas nécessaire ni même rationnel lorsque l'on considère des critères de vaisemblance pénalisée.

3.2.3 Les ratios de vraisemblance intégrée

Dans la logique de pondération bayésienne de modèles (BMA), de l'anglais "Bayesian Model Averaging" (cf. Madigan, Raftery, and Hoeting 1994 et Hoeting, Madigan, Raftery, and Volinsky 1999), les poids des modèles dans la combinaison sont estimés à partir de leurs vraisemblances intégrées. Hoeting, Madigan, Raftery, and Volinsky (1999) ont, par exemple, implémenté BMA pour des modèles de régression.

Selon BMA la combinaison de modèles s'exprime par

$$\mathbf{p}_{COM}(\mathbf{x}) = \sum_{m=1}^{M} \mathbf{p}_m(\mathbf{x})\beta_m \tag{3.22}$$

où $\mathbf{p}_{COM}(\mathbf{x})$ représente les probabilités a posteriori obtenues par combinaison et $\mathbf{p}_m(\mathbf{x})$ les probabilités a posteriori des groupes fournies par le modèle m.

Les coefficients de combinaison β_m ont dans ce contexte une signification spécifique. Ce sont les probabilités a posteriori $Pr(m \mid \mathbf{x})$ du modèle m. L'expression (3.22) est la moyenne des probabilités de chaque modèle pondéré par sa probabilité a posteriori.

La probabilité a posteriori du modèle m est calculée par la formule

$$\Pr(m \mid \mathbf{x}) = \frac{Pr(\mathbf{x}|m)Pr(m)}{\sum_{u=1}^{M} Pr(\mathbf{x}|u)Pr(u)} \tag{3.23}$$

où

$$Pr(\mathbf{x}|m) = L(\mathbf{x}|m) = \int Pr(\mathbf{x} \mid \theta_m)Pr(\theta_m)d\theta_m \tag{3.24}$$

est la vraisemblance intégrée ou marginale du modèle m,

$$Pr(\mathbf{x}|\theta_m) = L_m(\theta; \mathbf{x}) \tag{3.25}$$

étant la vraisemblance associée au modèle m et $Pr(\theta_m)$ la distribution a priori de θ_m. De plus, dans (3.23) $Pr(m)$ est la probabilité a priori que m soit le vrai modèle.

Ainsi, les coefficients β_m sont les ratios des vraisemblances intégrées

$$\tilde{\beta}_m = \frac{L(\mathbf{x}|m)}{\sum_{m=1}^{M} L(\mathbf{x}|m)}. \tag{3.26}$$

Malheureusement les intégrales de la forme de (3.24) sont difficiles à calculer[4]. Kass and Wasserman (1995), Raftery (1996) ont démontré que la vraisemblance intégrée peut être très bien approximée par le critère d'information bayésien (BIC) de Schwarz (1978).

$$BIC = -2\log L_m + \nu_m \log n. \tag{3.27}$$

[4] Pour les variables qualitatives il est possible d'avoir une formule explicite (Ferreira 2001).

D'où l'on déduit

$$e^{-\frac{BIC}{2}} = L_m(\theta; \mathbf{x}).n^{-0.5\nu_m}.$$

Et les coefficients β_m sont estimés par les ratios de vraisemblance intégrée aproximée où $L_m(\theta; \mathbf{x})$ est remplacée par $L_m(\theta; \mathbf{x}).n^{-0.5\nu_m}$

$$\beta_m = \frac{L_m(\theta; \mathbf{x}).n^{-0.5\nu_m}}{\sum_{u=1}^{M} L_u(\theta; \mathbf{x}).n^{-0.5\nu_u}}. \quad (3.28)$$

Finalement dans BMA et d'après la suggestion de Taplin (1993) les probabilités de chaque modèle $\mathbf{p}_m(\mathbf{x})$ sont estimés par l'approximation du maximum de vraisemblance. Les paramètres du modèle sont, ainsi, estimés par le maximum de vraisemblance. C'est ce que l'on a fait tout au long de ce travail.

Nous n'utiliserons pas la vraisemblance de validation croisée tout comme on ne l'a pas fait dans le cas de la vraisemblance pénalisée.

3.2.4 L'empilement

Wolpert (1992) est l'auteur de la proposition originale d'empilement, de l'anglais "stacking". En 1992, dans un contexte assez général il a proposé l'idée de "stacked generalisation" qui concerne l'utilisation de plusieurs modèles et le recueil d'information à deux niveaux. L'échantillon d'apprentissage constitue les données de niveau zéro et les échantillons obtenus par validation croisée les données de niveau un. En 1999 Smyth and Wolpert (1999) ont concrétisé cette forme combinée d'estimation de densités.

La combinaison de modèles par empilement vient de l'idée suivante. On suppose que l'échantillon d'apprentissage a été obtenu comme un mélange des K

groupes respectant les proportions $\pi_1, \ldots, \pi_K$. Chaque modèle propose une distribution de mélange pour les données de la forme

$$h_m(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_m^k(\mathbf{x}), \quad m = 1, \ldots, M. \tag{3.29}$$

Les π_k sont estimés par $\frac{n_k}{n}$ et les $f_m^k(\mathbf{x})$ sont estimés par maximum de vraisemblance avec utilisation de la validation croisée, comme précédemment.

Le principe de l'empilement consiste à déclarer que la densité des données est en fait un mélange des M distributions associées aux M modèles. C'est-à-dire que

$$f_{COM}(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}), \tag{3.30}$$

avec $\beta_m > 0$ pour tout m et $\sum_{m=1}^{M} \beta_m = 1$.

Maintenant les $h_m(\mathbf{x})$, estimés par validation croisée, sont supposés connus et on se retrouve avec un problème classique d'estimation des proportions d'un mélange de distributions connues. Plus précisément, on va chercher les β_m de sorte à maximiser la vraisemblance obtenue par validation croisée

$$\prod_{i=1}^{n} f_{COM}(\mathbf{x}_i) = \prod_{i=1}^{n} \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}_i)_{-i} \tag{3.31}$$

avec $h_m(\mathbf{x}_i)_{-i} = \sum_{k=1}^{K} \pi_k f_m^k(\mathbf{x}_i)_{-i}$. Cela se fait classiquement par l'algorithme EM que nous décrivons maintenant.

L'algorithme EM pour les poids, les composants étant fixes

Itération r

– étape E :

$$t_m^r(\mathbf{x}_i) = \frac{\beta_m^r f_m(\mathbf{x}_i)_{-i}}{\sum_{u=1}^{M} \beta_u^r f_m(\mathbf{x}_i)_{-i}}, \qquad m = 1, \ldots, M; \; i = 1, \ldots, n.$$

– étape M

$$\beta_m^{r+1} = \frac{\sum_{i=1}^{n} t_m^r(\mathbf{x}_i)}{n}.$$

Dans notre implantation, nous avons utilisé le critère d'arrêt suivant

$$\sum_{m=1}^{M} \mid \beta_m^{r+1} - \beta_m^r \mid < 0.0001.$$

3.3 La combinaison hiérarchique

Ce type de combinaison est de tout autre nature et s'applique lorsque le nombre de groupes à discriminer est plus grand que 2. La méthodologie de cette section est entièrement nouvelle.

Notre démarche consiste à utiliser une méthode de couplage des groupes à discriminer à chaque niveau d'une hiérarchie où la combinaison consiste à emboîter les modèles.

3.3.1 Le couplage hiérarchique

L'inspiration de cette méthode vient d'une idée de Friedman (1996). Il remarque que les règles de classification pour un problème à K groupes se révèlent plus faciles à interpréter et même à estimer dans le cas de $K = 2$ que dans le cas de $K > 2$. Ainsi, il a proposé l'approche suivante :

- construire les $\binom{K}{2}$ paires de groupes possibles à partir des K groupes,
- résoudre les $\binom{K}{2}$ problèmes de classification dichotomique,
- combiner toutes ces décisions afin d'avoir une décision pour le problème multigroupe.

Cette combinaison se fait pour Friedman par un vote à la majorité : l'observation à classer doit être attribuée à la classe qui aura gagné le plus de décisions dichotomiques.

En alternative à la stratégie du vote à la majorité, Hastie and Tibshirani (1996) ont aussi utilisé le couplage en analyse discriminante. Leur démarche consiste à

- construire les $\binom{K}{2}$ paires de groupes possibles à partir des K groupes,
- résoudre les $\binom{K}{2}$ problèmes de classification dichotomique,
- estimer les densités conditionnelles des K groupes à partir des densités conditionnelles dichotomiques estimées précédemment.

Pour cette dernière étape, les auteurs ont développé un modèle fondé sur un algorithme utilisé par les modèles log-linéaires, le "*iterative proportional scaling*" (Deming and Stephan 1940). Une fois obtenues les probabilités conditionnelles des K groupes, la règle de décision est la règle de Bayes habituel.

Notre approche est différente de ces deux approches. Nous allons, comme eux, décomposer un problème à plusieurs groupes en plusieurs problèmes à deux groupes. Mais les problèmes dichotomiques sont établis d'une autre façon ainsi que la régle de décision finale. Ces problèmes dichotomiques se trouvent, non pas indépendants mais, emboîtés dans une structure hiérarchique binaire, ce qui a pour premier effet de limiter le nombre de comparaisons à réaliser. Partant de là, toute observation à classer est affecté au groupe déterminé par la descente de l'arbre selon la règle de décision choisie à chaque nœud. On désignera notre méthode sous le nom de couplage hiérarchique.

La motivation pour cette technique est d'allier la simplicité d'estimation du problème dichotomique à la stabilité d'une combinaison de modèles. Cette stabilité vient du fait que l'affectation des points à l'un des K groupes n'est pas nécessairement attaché à un modèle unique.

3.3.2 La procédure

Soit $\mathcal{G}$ l'ensemble des K groupes $\mathcal{G} = \{G_1,\ G_2\ , ..., G_K\}$ et considérons une partition binaire de $\mathcal{G}$ appelée partition de premier niveau $P_1 = \{C_1^1,\ C_1^2\}$ où $C_1^1,\ C_1^2 \ \subset \mathcal{G}$, $C_1^1 \bigcup C_1^2 = \mathcal{G}$, $\ C_1^1 \bigcap C_1^2 = \emptyset$. Si par exemple $K = 4$, des partitions possibles sont $\{G_1,\ \{G_2, G_3, G_4\}\}, \{G_2,\ \{G_1, G_3, G_4\}\}$ ou $\{\{G_1, G_2\},\ \{G_3, G_4\}\}$.

À ce premier niveau les partitions possibles sont en nombre de $2^{K-1} - 1$. Chaque élément $C_1^l,\ l = 1, 2$ de la partition P_1 est nommé classe de premier niveau. Les classes qui ont un seul groupe sont les classes simples. Celles qui contiennent plus d'un groupe s'appellent classes composées.

Une règle de décision est appliquée au premier niveau afin de choisir la meilleure partition P_1^*. Les classes composées de P_1^* sont partitionnées, à leur tour, en deux classes dans ce qu'on appelle partition de deuxième niveau P_2. De nouveau la règle de décision intervient pour choisir la meilleure partition de deuxième niveau P_2^*. Et ainsi de suite jusqu'à ce que les classes de la partition courante soient seulement du type simple.

Cette procédure est représentée sous la forme d'un arbre binaire dont le nombre de branches est K. Ainsi pour $K = 4$ un arbre possible est représenté dans la figure 3.1.

3.3.3 La règle de décision

À chaque niveau de l'arbre nous disposons d'une paire de classes C_1 et C_2 sur lesquelles nous construisons une règle de discrimination dichotomique. Dans

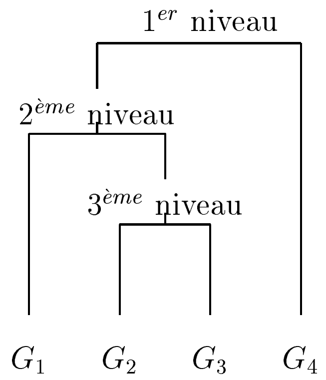


Figure 3.1 : Exemple d'un arbre possible représentatif du couplage hiérarchique pour un problème de discrimination à 4 groupes.

cette intention, les données sont supposées provenir d'un mélange à deux composants de populations gaussiennes. Les classes composées sont supposées être originaires du même composant du mélange. Le problème dichotomique est résolu avec les modèles EDDA.

Le couplage hiérarchique exige deux décisions à chaque niveau. L'une porte sur la bonne partition à choisir parmi toutes celles possibles. L'autre décision concerne le choix du modèle, parmi les M, qui discrimine le mieux les deux classes. Ce couplage de décision est construit de la façon suivante,

1. Pour toutes les partitions possibles P_1 estimer tous les modèles EDDA. Au premier niveau on aura $14.(2^{K-1} - 1)$ paires du type (modèle, partition) ;
2. Le choix désigne la paire avec le plus petit taux d'erreur de validation croisée. S'il y a plus qu'une paire (modèle, partition) avec le même taux d'erreur on choisira celle dont le modèle est le plus parcimonieux selon la règle de priorité définie en section 1.6.
3. Une fois obtenu la paire (m*, P_1^*) on passe au deuxième niveau et ainsi de suite.

Cette procédure est répétée pour tous les niveaux de l'arbre, aboutissant à une combinaison hiérarchique de modèles. Le modèle obtenu par combinaison sera aussi évalué par validation croisée.

Finalement le résultat d'une combinaison de modèles par couplage hiérarchique est reproduit dans la figure 3.2.

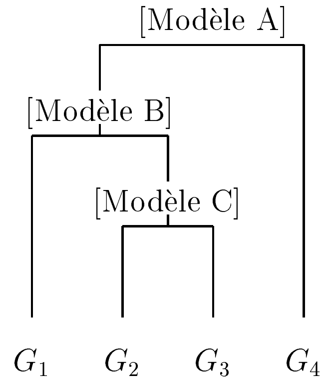


Figure 3.2 : Exemple de couplage hiérarchique.

Quand une nouvelle observation arrive elle passe par le modèle A qui la classe en G_4 ou en $G_1 \bigcup G_2 \bigcup G_3$. S'il s'agit du groupe G_4 l'analyse s'arrête. Autrement l'observation passe par le modèle B et la décision est le groupe G_1 ou $G_2 \bigcup G_3$. Si le modèle B ne classe pas en G_1 alors l'observation passe par le modèle C qui la classera en G_2 ou en G_3.

Remarque : le fait de choisir à chaque niveau de la hiérarchie la méthode et la partition des groupes qui fournissent le plus petit taux d'erreur ne garantit pas que l'arbre obtenu sera optimale. Nous avons expérimenté de nombreux cas où démarrer avec un couplage sous optimale conduisait finalement à un arbre meilleur que celui que l'on vient de décrire. Cependant, l'obtention de l'arbre optimale demande très vite un nombre de comparaisons énorme impraticable dès que $K \geq 4$. Il n'y a guère que pour $K = 3$ que l'on peut envisager d'explorer tous les arbres dichotomiques possibles. Notre solution, si elle n'est pas optimale à tout coup, nous semble en mesure de fournir des arbres de combinaison de modèles stables et de bonne qualité.

Chapitre 4

Performance des méthodes de combinaison de modèles

L'heure de la vérité.

4.1 Performance des différents méthodes de combinaison

Dans ce chapitre nous analysons la performance des méthodes de combinaison de modèles présentée dans le chapitre précédent. Dans le domaine de la combinaison de modèles, les résultats théoriques notamment asymptotiques sont inexistants et à vrai dire seraient de peu d'utilité car ils exigent des hypothèses irréalistes qui sont précisément contestées par la combinaison. Aussi une grande partie de notre travail a consisté à mener des expérimentations pour juger du bien-fondé des techniques de combinaison proposées, de leurs mérites comparés et de leurs principales caractéristiques par rapport au choix d'un unique modèle.

Deux sections distinctes s'occupent l'une des méthodes de combinaison en série (section 4.3) et l'autre de la combinaison hiérarchique (section 4.5). La performance de ces méthodes est évaluée sur des jeux de données réelles qui sont décrits dans la suite (section 4.2). Les méthodes de combinaison ont été construites avec les modèles gaussiens sous-jacents à EDDA. Par ailleurs, nous sommes sortis du cadre exclusivement gaussien en faisant une incursion non paramétrique (section 4.4). Finalement une comparaison de toutes les techniques de combinaison présentées sur chaque jeu de données synthétisant les expérimentations est présentée (section 4.6). Nous menons aussi une application sur des données de reconnaissance de tumeurs cérébrales, fournis par l'hôpital universitaire de Grenoble (section 4.7). Tous les méthodes de combinaison ont été programmées à l'aide du logiciel Matlab©.

4.2 Description des jeux de données

Pour ces expérimentations, nous avons utilisé 5 jeux de données originaires du "Machine Learning Repository" de l'Université de Californie (Blake and Merz 1998), un jeu de données tiré du dépôt de l'Université d'Oxford (Venables and Ripley 1994) et un autre jeu de données publié par Habbema, Hermans, and van den Broek (1974). Avant de les décrire, notons que chacun sera notée par un nom de référence avec lequel il est mentionné tout au long du texte. Leur origine est marquée, respectivement, par Cal., Oxf. et Hab.

bupa (Cal.) - Les données ont été assemblés par le groupe de recherche médical BUPA de Nottingham. Elles décrivent la présence ou l'absence de maladies du foie causées par une consommation excessive d'alcool. Pour 345 individus du sexe masculin 6 variables ont été observées : 5 sont les résultats de tests sanguins et la dernière est la quantité d'alcool prise par jour.

crabes (Oxf.) - Les observations en nombre de 200 portent sur la morphologie des crabes de la variété *Leptograpsus* qui se trouve en Australie. Les variables sont 5 différentes mesures des carapaces de deux spécimens de crabes : les oranges et les bleus. Il faut distinguer entre 4 groupes correspondant au sexe par spécimen.

haberman (Cal.) - Le jeu de données contient 306 observations d'une étude menée de 1958 à 1970 à l'hôpital universitaire Chicago Billings. L'étude concerne la survie de femmes qui ont pratiqué une chirurgie du sein pour ablation d'un cancer. Il y a deux groupes : si la patiente est décédée dans une période de cinq ans après l'opération ou si la patiente a survécu cinq ans ou plus. Les 3 variables explicatives sont l'âge du malade, l'année de la chirurgie et le nombre de zones cancérigènes détectés.

haemo (Hab.) - Un échantillon de 75 femmes classifiées selon qu'elles sont porteuses ou non d'hémophilie(du type A). Les femmes sont caractérisées par 2 variables médicales : l'activité AHF et l'antigène AHF.

iris (Cal.) - Ce sont les bien connues données de Fisher : 150 fleurs sont décrites par 4 variables, la longueur et la largeur de leurs sépale et pétale. Les fleurs appartiennent à 3 types différents d'iris : *setosa*, *versicolor* et *virginica*.

thyroïde (Cal.) - Des tests de laboratoire sont pratiqués sur des individus pour prédire l'état de la thyroïde : normal, hyper ou hypo fonctionement. Nous disposons de 215 observations de malades décrits par 5 variables de type médical.

vin (Cal.) - Les données sont les résultats d'analyses chimiques effectuées sur des vins originaires de la même région en Italie mais produits par trois cultivateurs différents. Les analyses ont déterminé les quantités de 13 éléments trouvés dans chacun des trois types de vins. Le jeu de données rassemble 178 observations, réparties sur trois groupes et caractérisées par 13 variables.

Le tableau 4.1 récapitule les jeux de données en termes du nombre d'individus de l'échantillon, du nombre de caractéristiques décrivant la population et du nombre de groupes.

Jeu de données	individus	caractéristiques	groupes
bupa	345	6	2
crabes	200	5	4
haberman	306	3	2
haemo	75	2	2
iris	150	4	3
thyroïde	215	5	3
vin	178	13	3

Tableau 4.1 : Nombre d'individus, de caractéristiques et de groupes pour les jeux de données référenciés.

Pour un aperçu descriptif de la possibilité de discriminer les classes à l'aide des données disponibles, nous avons effectué une analyse factorielle discriminante pour chacun des jeux de données excepté pour **haemo** qui n'est décrit que par 2 variables. La figure 4.1 donne la représentation des données dans le premier plan discriminant pour les données à plus de deux groupes. Encore dans la figure 4.1, pour les données à deux groupes, on donne leur histogramme sur l'unique axe discriminante sauf pour **haemo** où on donne leur description complète par les 2 variables qui caractérisent ces données.

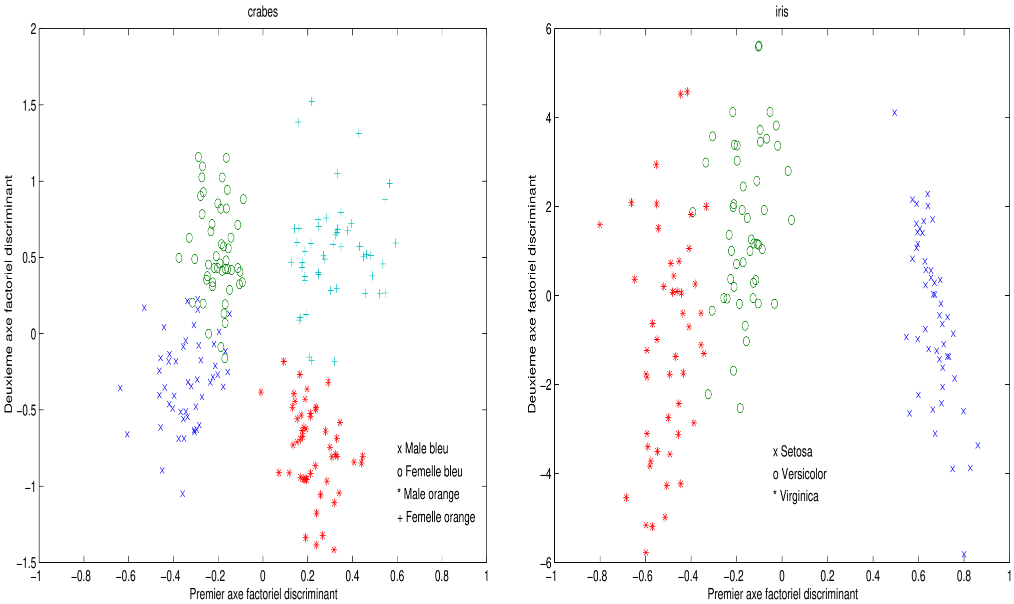


Figure 4.1(a) : Représentation de **crabes** et **iris** sur leurs deux premiers axes factoriels discriminants.

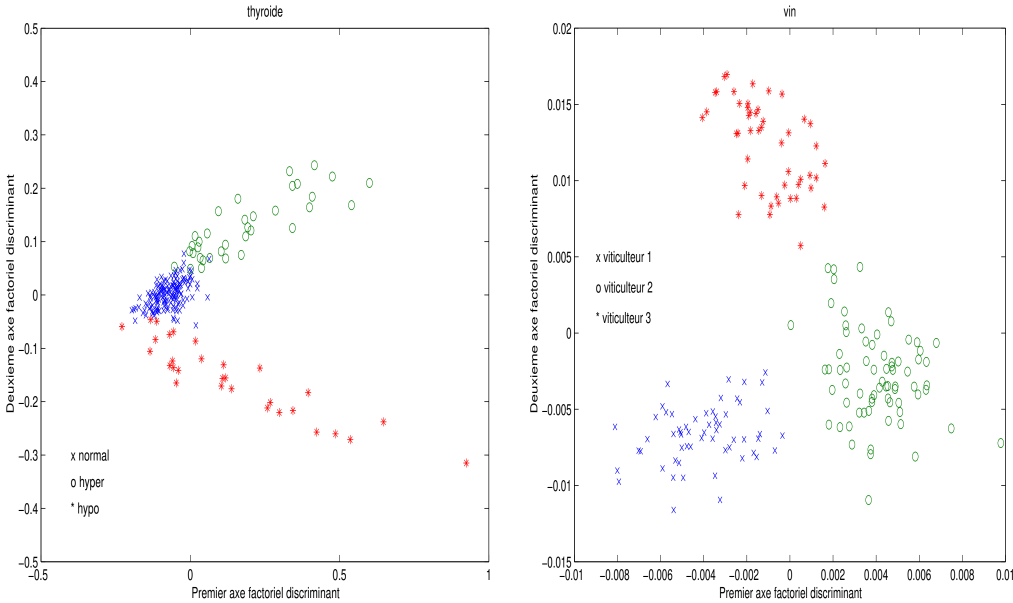


Figure 4.1(b) : Représentation de **thyroïde** et **vin** sur leurs deux premiers axes factoriels discriminants.

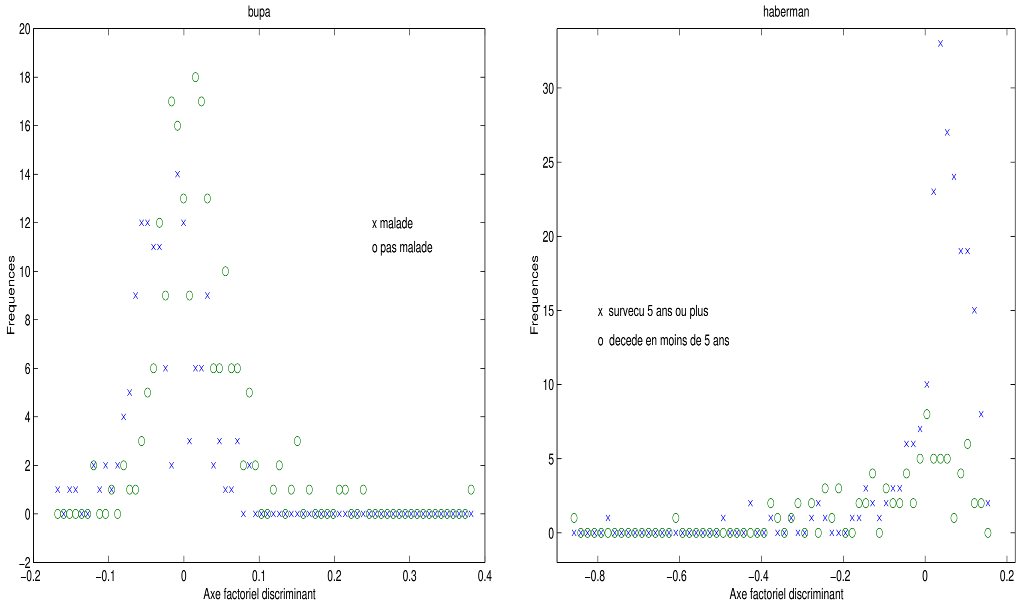


Figure 4.1(c) : Histogrames de **bupa** et **haberman** sur leur axe factoriel discriminant.

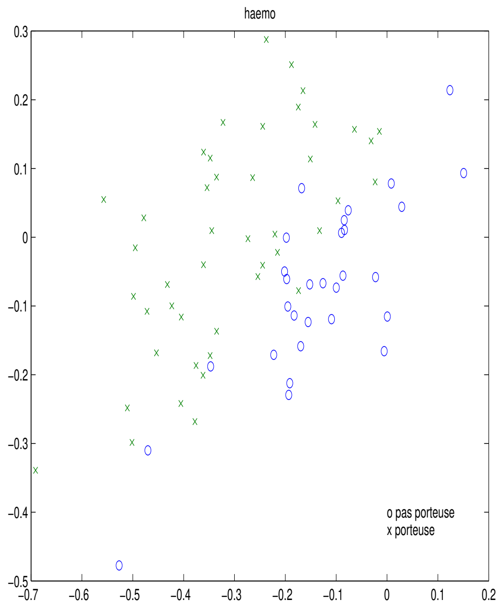


Figure 4.1(d) : Représentation de **haemo** sur ses variables d'origine.

À noter que pour certains jeux de données (**bupa** et **haberman**) les classes se présentent assez mélangées et pour d'autres (**crabes**, **iris**, **haemo**, **thyroïde** et **vin**), au contraire, les classes sont visiblement séparées. C'est cette dualité qui a motivé notre choix de ces jeux de données comme exemples de référence.

4.3 Performance des méthodes de combinaison en série

Nous avons expérimenté les techniques de combinaison envisagées dans le chapitre 3 pour combiner les quatorze modèles gaussiens comparés dans la méthode EDDA. Ce sont (cf. chapitre 1) :

1er type	$[\lambda\mathbf{DAD}']$, $[\lambda_k\mathbf{DAD}']$, $[\lambda\mathbf{DA}_k\mathbf{D}']$, $[\lambda_k\mathbf{DA}_k\mathbf{D}']$
	$[\lambda\mathbf{D}_k\mathbf{AD}_k']$, $[\lambda_k\mathbf{D}_k\mathbf{AD}_k']$, $[\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$, $[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$
2ème type	$[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{B}_k]$, $[\lambda_k\mathbf{B}_k]$
3ème type	$[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$

Le fait que beaucoup de ces quatorze modèles fournissent des règles de décision analogues, peut limiter l'intérêt de la combinaison de tous ces modèles. Aussi, dans un souci de mettre en évidence les gains possibles de la combinaison de modèles, nous avons également effectué des expérimentations en nous limitant aux trois modèles d'esprit le plus différent parmi les quatorze, à savoir l'analyse discriminante quadratique $[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$, l'analyse discriminante linéaire $[\lambda\mathbf{DAD}']$ et le classement à la moyenne la plus proche $[\lambda\mathbf{I}]$. Les sections concernées par cette stratégie de combinaison s'appelent "Combinaison avec les modèles les plus différents (I)".

De plus dans un même ordre d'idée, nous avons mis au point une stratégie qui, préalablement à la combinaison, sélectionne les modèles les plus éloignés de la façon suivante : on effectue une analyse de correspondances des probabilités par groupes fournies par chacun des modèles $p_m^k(\mathbf{x})$, $(m = 1, \ldots, 14)$. Puis la méthode des centres mobiles est utilisée (cf. par exemple Kaufman and Rousseeuw 1990) sur les coordonnées factoriels de tous les facteurs. Cette stratégie revient à utiliser l'algorithme des centres mobiles avec la distance du χ^2 (cf. par exemple Lebart, Morineau, and Piron 1995).

Le nombre de classes est fixé par inspection du dendograme. L'arbre est coupé où la distance entre deux niveaux est plus grande. Ensuite, de chaque classe nous avons pris le modèle le plus parcimonieux en nombre de paramètres (cf. chapitre 1). Dans le tableau 4.2 sont indiqués les modèles retenus pour chaque jeu de données avec cette stratégie.

Jeu de données	modèles considérés
bupa	$[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$
crabes	$[\lambda\mathbf{DAD}']$, $[\lambda\mathbf{I}]$
haberman	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$, $[\lambda\mathbf{B}_k]$, $[\lambda\mathbf{I}]$
haemo	$[\lambda\mathbf{DAD}']$, $[\lambda_k\mathbf{DAD}']$, $[\lambda\mathbf{I}]$
iris	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$
thyroïde	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$
vin	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$

Tableau 4.2 : Modèles retenus après classification des probabilités par groupe.

Les sections concernées par cette stratégie s'appelent "Combinaison avec les modèles les plus différents (II)".

Nous évaluons les méthodes de combinaison en série sur les jeux de données mentionnés. L'instrument d'évaluation est le taux d'erreur de classement (TEC) obtenu par la validation croisée standard (*leave-one-out*). En pratique le TEC est défini par le rapport entre le nombre d'observations mal classées par le modèle ou la technique discriminante et le nombre total d'observations.

Il est important de préciser ici que nous utilisons deux fois la validation croisée : comme on l'a vu, elle est utilisée pour déterminer les poids optimaux de la combinaison, puis pour évaluer le TEC. À chaque fois, c'est une double validation croisée qui est effectuée.

Enfin, il est important d'avoir à l'esprit que l'évaluation des méthodes de combinaison présentées dans cette thèse se fait presque exclusivement avec les modèles en compétition dans EDDA. Il est important de savoir si par rapport à cette technique la combinaison est pertinente. Le tableau 4.3 montre pour chaque jeu de données le meilleur modèle selon la stratégie EDDA et son TEC associé.

Jeu de données	Modèle EDDA	TEC
bupa	$[\lambda \mathbf{DAD}']$	0.3043
crabes	$[\lambda \mathbf{D}_k \mathbf{AD}'_k]$	0.0450
haberman	$[\lambda \mathbf{DA}_k \mathbf{D}']$	0.2418
haemo	$[\lambda \mathbf{DAD}']$	0.1467
iris	$[\lambda \mathbf{DAD}']$	0.0200
thyroïde	$[\lambda_k \mathbf{B}]$	0.0326
vin	$[\lambda \mathbf{D}_k \mathbf{AD}'_k]$	0[1]

Tableau 4.3 : Modèle d'analyse discriminante choisi et taux d'erreur estimé pour chaque jeu de données selon la méthodologie EDDA.

[1]Il est vrai que pour ce jeu de données il est impossible d'obtenir un TEC plus petit ! Nous avons néanmoins décidé de l'inclure. Dans ce cas nous voulons analyser les modèles présents dans la combinaison.

Dans les sections concernant les stratégies de combinaison avec les modèles les plus différents (I) et (II) l'évaluation des méthodes de combinaison se fait par rapport au meilleur modèle parmi ceux qui sont considérés dans la combinaison. Les tableaux 4.4 et 4.5 montrent les meilleurs modèles pour chaque jeu de données et son TEC associé dans le cas de chaque stratégie.

Jeu de données	Modèle	TEC
bupa	$[\lambda \mathbf{DAD}']$	0.3043
crabes	$[\lambda \mathbf{DAD}']$	0.0500
haberman	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.2451
haemo	$[\lambda \mathbf{DAD}']$	0.1467
iris	$[\lambda \mathbf{DAD}']$	0.0200
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0372
vin	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0056

Tableau 4.4 : Modèle d'analyse discriminante choisi parmi les $[\lambda \mathbf{DAD}']$, $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$ et $[\lambda \mathbf{I}]$ et taux d'erreur estimé pour chaque jeu de données.

Jeu de données	Modèle	TEC
bupa	$[\lambda \mathbf{B}]$	0.4000
crabes	$[\lambda \mathbf{I}]$	0.0500
haberman	$[\lambda \mathbf{B}_k]$	0.2516
haemo	$[\lambda \mathbf{DAD}']$	0.1467
iris	$[\lambda \mathbf{B}]$	0.0400
thyroïde	$[\lambda \mathbf{B}]$	0.0977
vin	$[\lambda \mathbf{B}]$	0.0449

Tableau 4.5 : Modèle d'analyse discriminante choisi parmi les modèles retenues après classification des probabilités par groupe et taux d'erreur estimé pour chaque jeu de données.

Les sections suivantes consistent en une collection de tableaux qui présentent le TEC obtenu sur les différents jeux de données pour les techniques de combinaison en étude. Rapellons que les techniques que nous avons retenu sont,

- La régression séparée pour chaque groupe - (4.3.1) ;
- La régression vectorisée - (4.3.2) ;
- Le comité de méthode - (4.3.3) ;
- Le ratio de vraisemblance - (4.3.4) ;
- Le ratio de vraisemblance pénalisée - (4.3.5) ;
- Le ratio de vraisemblance intégrée - (4.3.6) ;
- L'empilement - (4.3.7) ;

Dans nos expérimentations nous avons trouvé les coefficients attribués à chaque modèle dans la combinaison $\tilde{\beta}_m$, $(m = 1, \ldots, M)$ par les différentes techniques de combinaison. Ensuite nous avons procédé à la validation de la technique en utilisant quelques variantes, à partir des poids obtenus $\tilde{\beta}_m$. Ainsi nous avons essayé la combinaison avec,

- les coefficients $\tilde{\beta}_m$ déduits de la contrainte de non negativité ($\beta_m \geq 0$) bruts,
- les coefficients $\tilde{\beta}_m$ supérieurs à 0.1 - l'idée étant de ne pas inclure les modèles qui contribuent peu à la combinaison -,
- les coefficients $\tilde{\beta}_m$ supérieurs à 0.1 et arrondis - l'idée étant de simplifier la combinaison - ,
- les coefficients $\tilde{\beta}_m$ supérieurs à 0.1 et normalisées ($\sum_{m=1}^{M}(\tilde{\beta}_m > 0.1) = 1$),
- les coefficients $\tilde{\beta}_m$ déduits d'une contrainte supplémentaire : de somme unitaire ($\sum_{m=1}^{M} \beta_m = 1$).

Toutes ces variations n'ont pas été considérées pour chaque méthode car il y a des situations où elles n'ont pas grand sens. Cela sera précisé dans la section adéquate.

Quand cela s'avère intéressant on présente dans ce chapitre les tableaux des poids $\tilde{\beta}_m$ attribués à chaque modèle dans la combinaison. Sinon cette information est reléguée à l'annexe où sont réunis tous les tableaux des poids pour toutes les méthodes de combinaison.

4.3.1 Régression séparée

Combinaison avec 14 modèles

Jeu de données	contrainte de non négativité					modèle EDDA
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3101	0.3043	0.3130	0.3101	0.3043	0.3043
crabes	0.0550	0.0550	0.0550	0.0550	0.0550	0.0450
haberman	0.2582	0.2582	0.2582	0.2582	0.2582	0.2418
haemo	0.1733	0.1733	0.1733	0.1600	0.1600	0.1467
iris	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200
thyroïde	0.0465	0.0465	0.0465	0.0465	0.0465	0.0326
vin	0.0056	0.0056	0.0056	0.0056	0.0056	0

Tableau 4.6 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression separée* avec les 14 modèles EDDA.

La combinaison par régression séparée fait dans deux cas, **bupa** et **iris**, aussi bien que le meilleur modèle EDDA. La variation autour des poids entraîne dans deux cas, **bupa** et **haemo**, une différence des TEC. Pour **bupa** c'est les $\tilde{\beta}_m > 0.1$ et la contrainte unitaire qui ont le plus petit TEC. Pour **haemo** c'est les $\tilde{\beta}_m > 0.1$ normalisés et la contrainte unitaire qui ont une meilleure performance. Un désavantage de cette technique de combinaison est l'existence d'un grand nombre de coefficients, ce qui rend son interprétation quelque peu difficile. Dans le tableau 4.7 nous présentons le modèle d'analyse discriminante qui résulte de la combinaison pour quelques jeux de données. En annexe les modèles pour tous les jeux de données sont présentés (tableau A.1).

Jeu de données	groupe	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	groupe 1	$0.39[\lambda\mathbf{DAD}'] + 0.50[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.11[\lambda\mathbf{I}]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.39[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.13[\lambda\mathbf{I}]$	
crabes	groupe 1	$0.17[\lambda\mathbf{DAD}'] + 0.14[\lambda_k\mathbf{DAD}']+$ $+0.3[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.39[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$
	groupe 2	$0.35[\lambda_k\mathbf{DAD}'] + 0.65[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 3	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 4	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
haemo	groupe 1	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
	groupe 2	$[\lambda\mathbf{DAD}']$	
iris	groupe 1	$[\lambda_k\mathbf{B}]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.3[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.7[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 3	$0.24[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.76[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	

Tableau 4.7 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les 14 modèles EDDA.

Pour chaque groupe considéré dans la régression, le modèle d'analyse discriminante se lit comme la position de ce groupe vis-à-vis de tous les autres groupes réunis. C'est ainsi que pour **bupa** - un jeu de données à deux groupes - les modèles composants de la combinaison sont les mêmes pour le groupe 1 que pour le groupe 2. Aussi les poids attribués à chaque modèle sont sensiblement les mêmes dans le cas des deux groupes. Pareille situation se vérifie pour **haemo**, mais en plus, nous pouvons déduire que la combinaison est superflue car il y a un seul modèle qui compte. En ce qui concerne **iris**, par exemple, le groupe 1 - les *setosa* - est discriminé des autres deux groupes par le modèle diagonale $[\lambda_k \mathbf{B}]$. Tandis que le groupe 2 - les *versicolor* - est discriminé par la combinaison des modèles $[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$ et $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ dont la participation du deuxième modèle est plus importante que celle du premier.

Combinaison avec les modèles les plus différents (I)

Les modèles considérés ici sont l'analyse discriminante quadratique $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$, l'analyse discriminante linéaire $[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$ et le classement à la moyenne la plus proche $[\lambda \mathbf{I}]$.

Jeu de données	contrainte de non négativité					modèle EDDA
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3188	0.3304	0.3304	0.3159	0.3188	0.3043
crabes	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
haberman	0.2516	0.2549	0.2549	0.2549	0.2516	0.2451
haemo	0.1733	0.1733	0.1733	0.1600	0.1600	0.1467
iris	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200
thyroïde	0.0372	0.0419	0.0419	0.0372	0.0372	0.0372
vin	0.0056	0.0056	0.0056	0.0056	0.0056	0.0056

Tableau 4.8 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression separée* avec les modèles les plus différents (I).

La combinaison avec ces trois modèles fait aussi bien que la stratégie EDDA à trois modèles dans le cas de **crabes**, **iris**, **tyroïde** et **vin**. Pour ces jeux de données, les variations autour des $\tilde{\beta}_m$ ne conduisent pas à des modifications de TEC excepté dans le cas de **tyroïde** où les $\tilde{\beta}_m > 0.1$ et les $\tilde{\beta}_m > 0.1$ arrondis sont moins bons. Les modèles composants la combinaison et les poids respectifs sont présentés au tableau A.2 de l'annexe.

Combinaison avec les modèles les plus différents (II)

Les modèles qui jouent maintenant dans la combinaison sont les modèles plus éloignés, ceux-ci étant déterminé par une classification des probabilités par groupe de chaque modèle. Les modèles retenus pour chaque jeu de données sont présentés dans le tableau 4.2.

De la lecture du tableau 4.9 on constate que la présente stratégie arrive à fournir des TEC plus petits que la stratégie EDDA dans le cas de **bupa** et de **tyroïd**. Les deux stratégies ont des performances égales pour **crabes**, **iris** et **vin**. Il s'avère que, pour la méthode de combinaison régression séparée, l'actuelle stratégie est la plus attractive. Ce cas montre qu'il est possible d'obtenir des taux d'erreur plus faibles par la combinaison de méthodes.

Jeu de données	contrainte de non négativité					modèle EDDA
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3884	0.3884	0.3913	0.3971	0.3971	0.4000
crabes	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
haberman	0.2549	0.2582	0.2582	0.2549	0.2549	0.2516
haemo	0.1733	0.1733	0.1733	0.1600	0.1600	0.1467
iris	0.0400	0.0400	0.0400	0.0400	0.0400	0.0400
thyroïde	0.0884	0.0884	0.0884	0.0930	0.0930	0.0977
vin	0.0449	0.0449	0.0449	0.0449	0.0449	0.0449

Tableau 4.9 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression separée* avec les modèles les plus différents (II).

4.3.2 Régression vectorisée

Combinaison avec 14 modèles

Jeu de données	contrainte de non négativité					modèle
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	modèle EDDA
bupa	0.3159	0.3275	0.3304	0.3275	0.3420	0.3043
crabes	0.0550	0.0550	0.0550	0.0550	0.0650	0.0450
haberman	0.2614	0.2582	0.2582	0.2484	0.3497	0.2418
haemo	0.1733	0.1733	0.1733	0.1600	0.1600	0.1467
iris	0.0200	0.0200	0.0200	0.0200	0.6400	0.0200
thyroïde	0.0419	0.0465	0.0465	0.0465	0.1302	0.0326
vin	0.0056	0.0056	0.0056	0.0056	0.4551	0

Tableau 4.10 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression vectorisée* avec les 14 modèles EDDA.

Par rapport à la combinaison par régression séparée (tableau 4.6) seul le TEC de **haberman** pour les $\tilde{\beta}_m > 0.1$ normalisés est plus petit (mais il est quand même supérieur à celui d'EDDA). Dans les autres cas le TEC est égal ou plus grand. Cette stratégie de combinaison produit beaucoup trop de coefficients tel qu'on peut le voir sur le tableau 4.11. Le modèle d'analyse discriminante devient, alors, ininterpretable. La lecture à faire de ce modèle est la suivante. Pour **crabes** par exemple, le groupe 1 - les mâles bleus - est séparé des autres groupes par le modèle combiné $.33[\lambda \mathbf{DAD'}] + .28[\lambda_k \mathbf{DA}_k\mathbf{D'}] + .38[\lambda \mathbf{D}_k\mathbf{AD}'_k]$ qui est naturellement différent de celui trouvé par la régression séparée.

Jeu de données	modèles contribuant à la combinaison et poids respectifs				modèle EDDA
crabes groupe régressé	groupe 1	groupe 2	groupe 3	groupe 4	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
groupe 1	$.33[\lambda \mathbf{DAD}']$ $+.28[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$ $+.38[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$				
groupe 2		$.35[\lambda_k \mathbf{DAD}']$ $+.65[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$			
groupe 3			$[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$		
groupe 4				$[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$	
haemo groupe régressé	groupe 1	groupe 2			$[\lambda \mathbf{DAD}']$
groupe 1	$[\lambda \mathbf{DAD}']$				
groupe 2		$[\lambda \mathbf{DAD}']$			
haberman groupe régressé	groupe 1	groupe 2			$[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$
groupe 1	$.38[\lambda \mathbf{DAD}']$ $+.20[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ $+.20[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$+.22[\lambda_k \mathbf{I}]$			
groupe 2		$.46[\lambda \mathbf{DAD}']$ $+.32[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ $+.22[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			

Tableau 4.11 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données pour la méthode de combinaison *régression vectorisée* avec les 14 modèles EDDA.

Pour **haemo** seul le modèle linéaire discrimine au mieux les deux groupes, indicant que la combinaison de modèles est une technique superflue dans ce cas. Plus difficile à interpréter est le modèle qui dérive de la combinaison pour **haberman**. Le groupe 1 - le groupe des patients qui ont survécu 5 ans ou plus - est discriminé par un modèle qui comporte un coefficient concernant le groupe 2 - le groupe des patients qui ont décédé en moins de 5 ans - avec un poids d'à peu près 20%. Le tableau des modèles pour tous les jeux de données est en annexe (tableau A.4).

Combinaison avec les modèles les plus différents (I)

Jeu de données	contrainte de non négativité					modèle EDDA
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3188	0.3304	0.3304	0.3159	0.4406	0.3043
crabes	0.0550	0.0500	0.0500	0.0500	0.0800	0.0500
haberman	0.2582	0.2582	0.2582	0.2451	0.3170	0.2451
haemo	0.1733	0.1600	0.1733	0.1600	0.1600	0.1467
iris	0.0400	0.0400	0.0400	0.0400	0.3000	0.0200
thyroïde	0.0372	0.0372	0.0372	0.0372	0.2279	0.0372
vin	0.0056	0.0056	0.0056	0.0056	0.3371	0.0056

Tableau 4.12 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (I).

En combinant les modèles d'analyse discriminante quadratique $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$, linéaire $[\lambda \mathbf{DAD}']$ et le classement à la moyenne la plus proche $[\lambda \mathbf{I}]$, on obtient un TEC égale au meilleur de ces modèles pour **crabes**, **haberman**, **thyroïde** et **vin**. Pour ces jeux de données le plus petit TEC parvient de la considération des $\tilde{\beta}_m > 0.1$ normalisés dans tous les cas, des $\tilde{\beta}_m > 0.1$ et $\tilde{\beta}_m > 0.1$ arrondis dans trois cas et des $\tilde{\beta}_m$ dans deux cas. Le tableau A.5 de l'annexe présente les modèles contribuant à la combinaison et les poids respectifs pour chaque jeu de données.

Combinaison avec les modèles les plus différents (II)

Jeu de données	contrainte de non négativité					modèle EDDA
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3913	0.3942	0.3942	0.3797	0.4899	0.4000
crabes	0.0550	0.0500	0.0500	0.0500	0.0750	0.0500
haberman	0.2614	0.2647	0.2647	0.2549	0.3301	0.2516
haemo	0.1733	0.1600	0.1733	0.1600	0.1867	0.1467
iris	0.0400	0.0400	0.0400	0.0400	0.3000	0.0400
thyroïde	0.0884	0.0884	0.0884	0.0930	0.0930	0.0977
vin	0.0056	0.0056	0.0056	0.0056	0.0955	0.0449

Tableau 4.13 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (II).

De même que pour la méthode de régression séparée, la présente stratégie parvient à des TEC plus petits que celui du meilleur modèle dans trois cas : **bupa**, **thyroïde** et **vin**. Il est encore possible de faire aussi bien pour les deux cas **crabes** et **iris**. On remarque que la méthode de régression vectorisée accuse une baisse notable et curieuse des résultats associés à la contrainte unitaire. Les modèles contribuant à la combinaison et les pondérations associées sont montrés au tableau A.6 de l'annexe.

4.3.3 Comité de méthodes

Combinaison avec 14 modèles

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3043	0.3043
crabes	0.0500	0.0450
haberman	0.2582	0.2418
haemo	0.1600	0.1467
iris	0.0200	0.0200
thyroïde	0.0372	0.0326
vin	0.0056	0

Tableau 4.14 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *comité de méthodes* avec les 14 modèles EDDA.

La considération de variations sur les $\tilde{\beta}_m$ ne se pose pas car, pour cette technique, on est toujours dans le cas de la contrainte de somme unitaire des poids attribués à chaque modèle dans la combinaison. Les TEC obtenus par cette méthode sont au niveau des précédents TEC obtenus par les méthodes de régression séparée et vectorisée. Dans deux cas, **bupa** et **iris** le comité de méthodes fait aussi bien que le meilleur modèle EDDA. Un atout de la présente technique est le fait que le modèle d'analyse discriminante présente moins de coefficients comme on peut le voir dans le tableau 4.15. La technique comité de méthodes devient, de ce fait, plus attractive.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$0.41[\lambda \mathbf{DAD}'] + 0.48[\lambda \mathbf{D}_k \mathbf{AD}'_k] + 0.11[\lambda \mathbf{I}]$	$[\lambda \mathbf{DAD}']$
crabes	$0.16[\lambda_k \mathbf{DAD}'] + 0.83[\lambda \mathbf{D}_k \mathbf{AD}'_k]$	$[\lambda \mathbf{D}_k \mathbf{AD}'_k]$
haemo	$[\lambda \mathbf{DAD}']$	$[\lambda \mathbf{DAD}']$

Tableau 4.15 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données pour la méthode de combinaison *comité de méthodes* avec les 14 modèles EDDA.

Par exemple, le jeu de données **crabes** discrimine ses quatre groupes au moyen d'un modèle qui est composé en 16% du modèle $[\lambda_k \mathbf{DAD}']$ - les densités des groupes diffèrent en volume - et en 83% du modèle $[\lambda \mathbf{D}_k \mathbf{AD}'_k]$ - les densités des groupes diffèrent en orientation -. Aussi le comité de méthodes ne prend qu'un modèle, le linéaire, dans le cas de **haemo**. Les modèles d'analyse discriminante obtenus par le comité de méthodes avec les quatorze modèles EDDA pour tous les jeux de données sont présentés au tableau A.7 de l'annexe.

Combinaison avec les modèles les plus différents (I)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3188	0.3043
crabes	0.0550	0.0500
haberman	0.2516	0.2451
haemo	0.1600	0.1467
iris	0.0200	0.0200
thyroïde	0.0419	0.0372
vin	0.0112	0.0056

Tableau 4.16 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *comité de méthodes* avec les modèles les plus différents (I).

Nous constatons que la stratégie de combinaison fondée sur les modèles linéaire, quadratique et le classement à la moyenne la plus proche appliquée au comité de méthodes en aucun cas, parmi les cas étudiés, ne fait aussi bien que le meilleur modèle EDDA. Les modèles contribuant à la combinaison et les poids respectifs sont présentés au tableau A.8 de l'annexe.

Combinaison avec les modèles les plus différents (II)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3971	0.4000
crabes	0.0500	0.0500
haberman	0.2549	0.2516
haemo	0.1600	0.1467
iris	0.0400	0.0400
thyroïde	0.0930	0.0977
vin	0.0449	0.0449

Tableau 4.17 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *comité de méthodes* avec les modèles les plus différents (II).

Le TEC associée à la présente stratégie est légèrement plus petit que celui du meilleur modèle EDDA parmi ceux qui ont été retenus par l'analyse de correspondances, pour **bupa** et **thyroïde**. Les TEC sont pareils pour **crabes**, **iris** et **vin**. Les modèles composants de la conbinaison sont donnés dans le tableau A.9 de l'annexe.

4.3.4 Ratios des vraisemblances

Rappelons que les techniques présentées dans cette section et dans les prochaines quatre sections sont d'esprit différent des techniques précédentes car elles s'appuient sur la vraisemblance des modèles à combiner. Nous considérons toujours les trois stratégies de combinaison : avec les quatorze modèles, avec les modèles linéaire, quadratique et classement à la moyenne la plus proche et avec les modèles retenus par classification sur la base.

Les deux contraintes de non négativité et de somme unitaire des poids sont l'apanage des méthodes des ratios des vraisemblances, que ce soit pour les vraisemblances pures, pénalisées ou intégrées. Ainsi, on ne tient pas compte de variations sur les $\tilde{\beta}_m$ car on est toujours dans le cas de contrainte unitaire.

Combinaison avec 14 modèles

D'après le tableau 4.18, nous constatons que les TEC associés à cette méthode de combinaison dépassent dans tous les cas les TEC associés au meilleur modèle EDDA. Cette stratégie s'avère ainsi, assez décevante. Le tableau 4.19 montre la principale limitation de cette technique.

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3855	0.3043
crabes	0.0650	0.0450
haberman	0.2484	0.2418
haemo	0.1600	0.1467
iris	0.0267	0.0200
thyroïde	0.0372	0.0326
vin	0.0056	0

Tableau 4.18 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios des vraisemblances* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}\mathbf{A}\mathbf{D}']$
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$
haberman	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}\mathbf{A}_k\mathbf{D}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}\mathbf{A}\mathbf{D}']$
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{B}']$

Tableau 4.19 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données pour la méthode de combinaison *ratios des vraisemblances* avec les 14 modèles EDDA.

La combinaison par les ratios de vraisemblance a le désavantage d'attribuer plus d'importance, c'est-à-dire, un grand coefficient, invariablement au modèle le plus complexe. Parmi les quatorze, le modèle qui a le plus de paramètres à estimer est le modèle quadratique. Pour les jeux de données présentées au tableau 4.19, le poids attribué au modèle quadratique est à sa valeur extrême : 1. Les modèles composants de la combinaison pour tous les jeux de données sont montrés au tableau A.10 de l'annexe.

Combinaison avec les modèles les plus différents (I)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4058	0.3043
crabes	0.0650	0.0500
haberman	0.2451	0.2451
haemo	0.1600	0.1467
iris	0.0267	0.0200
thyroïde	0.0372	0.0372
vin	0.0056	0.0056

Tableau 4.20 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios des vraisemblances* avec les modèles les plus différents (I).

Naturellement ce sont les jeux de données **haberman**, **thyroïde** et **vin** qui exhibent un TEC comparable à celui du meilleur modèle EDDA. En effet, pour ces jeux de données le modèle quadratique est le meilleur pour EDDA parmi ceux qui sont présents dans la combinaison (cf. tableau 4.4) tel qu'on peut le voir sur le tableau A.11 de l'annexe.

Combinaison avec les modèles les plus différents (II)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4290	0.4000
crabes	0.0500	0.0500
haberman	0.2516	0.2516
haemo	0.1600	0.1467
iris	0.0400	0.0400
thyroïde	0.0977	0.0977
vin	0.0449	0.0449

Tableau 4.21 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios de vraisemblance* avec les modèles les plus différents (II).

La grande limitation de la présente technique est aussi visible sur cette stratégie, ce qui lui enlève beaucoup de son intérêt. Dans le tableau A.12 de l'annexe les modèles contribuant à la combinaison par cette stratégie sont donnés.

4.3.5 Ratios des vraisemblances pénalisées

Pour corriger la dépendance de la technique des ratios des vraisemblances à la compléxité des modèles nous avons introduit une pénalisation. Cette pénalisation est fondée sur le nombre de paramètres du modèle et dérive du critère AIC.

Combinaison avec 14 modèles

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3884	0.3043
crabes	0.0450	0.0450
haberman	0.2516	0.2418
haemo	0.1467	0.1467
iris	0.0267	0.0200
thyroïde	0.0372	0.0326
vin	0.0056	0

Tableau 4.22 : Taux d'erreur de classement pour chaque jeu de données par la méthode de combinaison *ratios de vraisemblance penalisée* avec les 14 modèles EDDA.

Les ratios des vraisemblances pénalisées produisent pour **crabes** et **haemo** des TEC au niveau du meilleur modèle EDDA. Tel qu'on peut le voir au tableau 4.23, pour **crabes** cette méthode de combinaison propose un seul modèle ce qui est peu intéressant. Par contre pour **haemo** plusieurs modèles sont pris en compte dans la combinaison.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
haemo	$.41[\lambda \mathbf{DAD}'] + .3[\lambda_k \mathbf{DAD}'] + .3[\lambda \mathbf{D}\mathbf{A}_k\mathbf{D}']$ $+.14[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k] + .39[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$

Tableau 4.23 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances pénalisées* avec les 14 modèles EDDA.

Combinaison avec les modèles les plus différents (I)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3043	0.3043
crabes	0.0650	0.0500
haberman	0.2516	0.2451
haemo	0.1600	0.1467
iris	0.0267	0.0200
thyroïde	0.0372	0.0372
vin	0.0112	0.0056

Tableau 4.24 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios de vraisemblance penalisée* avec les modèles les plus différents (I).

Par cette stratégie **bupa** et **thyroïde** ont des TEC comparables aux TEC associés au meilleur EDDA présents dans la combinaison. Pourtant dans les deux cas, seul un modèle résulte de la combinaison : pour **bupa**, le modèle linéaire et pour **thyroïde** le modèle quadratique (cf. tableau A.14 de l'annexe).

Combinaison avec les modèles les plus différents (II)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4000	0.4000
crabes	0.0500	0.0500
haberman	0.2516	0.2516
haemo	0.1467	0.1467
iris	0.0400	0.0400
thyroïde	0.0977	0.0977
vin	0.0449	0.0449

Tableau 4.25 : Taux d'erreur de classement pour chaque jeu de données par la méthode de combinaison *ratios de vraisemblance penalisée* avec les modèles les plus différents (II).

Par cette stratégie la méthode de combinaison ratios de vraisemblance pénalisée fait dans tous les cas aussi bien que le meilleur modèle EDDA présent dans la combinaison. Pourtant la conclusion est falacieuse car à l'exception de **haemo** pour tous les jeux de données il y a qu'un seul modèle contribuant à la combinaison (cf. tableau 4.15 de l'annexe). En tout cas, cela indique qu'il n'y a pas lieu d'attendre d'amélioration du taux d'erreur de classement par la combinaison de méthodes pour ces jeux de données.

Il se peut que l'usage de la validation croisée standard ne soit pas très heureux que l'utilisation de la validation par blocs puisse donner des résultats plus contrastés notamment pour de grandes tailles d'échantillons.

4.3.6 Ratios des vraisemblances intégrées

Dans la logique de pondération bayésienne de modèles, les poids des modèles dans la combinaison sont estimés à partir de leurs vraisemblances intégrées. Celle est très bien approximée par le critère BIC.

Combinaison avec 14 modèles

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3913	0.3043
crabes	0.0500	0.0450
haberman	0.2549	0.2418
haemo	0.1467	0.1467
iris	0.0267	0.0200
thyroïde	0.0326	0.0326
vin	0.0225	0

Tableau 4.26 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios de vraisemblance intégrée* avec les 14 modèles EDDA.

La méthode de combinaison fondée sur les vraisemblances intégrées fait aussi bien qu'EDDA pour **haemo** et **thyroïde** et moins bien dans les autres cas. Sur le tableau 4.27 on peut lire les modèles contribuant à la combinaison et les poids respectifs pour quelques jeux de données. Pour **thyroïde** il y a qu'un seul modèle qui compte ce qui limite l'intérêt de la combinaison.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
crabes	$.65[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k] + .35[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
haberman	$.42[\lambda \mathbf{B}_k] + .58[\lambda_k \mathbf{B}_k]$	$[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$
haemo	$.51[\lambda \mathbf{DAD}'] + .34[\lambda_k \mathbf{DAD}'] + .15[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{B}_k]$	$[\lambda_k \mathbf{B}]$

Tableau 4.27 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances intégrées* avec les 14 modèles EDDA.

Combinaison avec les modèles les plus différents (I)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3043	0.3043
crabes	0.0500	0.0500
haberman	0.2516	0.2451
haemo	0.1467	0.1467
iris	0.0200	0.0200
thyroïde	0.0372	0.0372
vin	0.0112	0.0056

Tableau 4.28 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios de vraisemblance intégrée* avec les modèles les plus différents (I).

À l'exception de **haberman** et **vin** dans tous les cas le TEC associé à cette stratégie est comparable au meilleur EDDA. Pourtant souvent la technique ne propose qu'un modèle (cf. tableau A.17 de l'annexe), à l'exception notable de **haberman** et **haemo**.

Combinaison avec les modèles les plus différents (II)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4000	0.4000
crabes	0.0500	0.0500
haberman	0.2516	0.2516
haemo	0.1467	0.1467
iris	0.0400	0.0400
thyroïde	0.0977	0.0977
vin	0.0449	0.0449

Tableau 4.29 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *ratios de vraisemblance intégrée* avec les modèles les plus différents (II).

Par cette stratégie la méthode de combinaison ratios de vraisemblance intégrée fait dans tous les cas aussi bien que le meilleur modèle EDDA présent dans la combinaison. Mais comme nous l'avions remarqué dans la section précédente la conclusion est falacieuse car à l'exception de **haemo** pour tous les jeux de données il y a qu'un seul modèle contribuant à la combinaison (cf. tableau A.18 de l'annexe) et le même commentaire que nous avon fait dans la section 4.3.5 peut être répété.

4.3.7 L'empilement

Le principe de l'empilement consiste à déclarer que la densité des données est un mélange des distributions associées aux modèles. Les poids attribués aux modèles dans la combinaison sont calculés de façon à maximiser la vraisemblance.

Combinaison avec 14 modèles

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.3855	0.3043
crabes	0.0550	0.0450
haberman	0.2516	0.2418
haemo	0.1600	0.1467
iris	0.0200	0.0200
thyroïde	0.0419	0.0326
vin	0.0056	0

Tableau 4.30 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *empilement* avec les 14 modèles EDDA.

L'empilement fait aussi bien que le meilleur modèle EDDA pour **iris** et moins bien dans tous les autres cas. Pour ce jeu de données le modèle d'analyse discriminante obtenu par combinaison est présenté au tableau 4.31.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
iris	$.2[\lambda_k \mathbf{DAD'}] + .15[\lambda \mathbf{D}_k \mathbf{AD}'_k] + .65[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD'}]$

Tableau 4.31 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *empilement* avec les 14 modèles EDDA.

Combinaison avec les modèles les plus différents (I)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4145	0.3043
crabes	0.0600	0.0500
haberman	0.2451	0.2451
haemo	0.1600	0.1467
iris	0.0200	0.0200
thyroïde	0.0372	0.0372
vin	0.0112	0.0056

Tableau 4.32 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *empilement* avec les modèles les plus différents (I).

L'empilement avec la stratégie des modèles les plus différents (I) produit des TEC égaux aux TEC des meilleurs modèles EDDA pour **haberman**, **iris** et **thyroïde**. Mais cette performance repose sur le fait qu'un seul un modèle est pris en compte (cf. tableau A.20 de l'annexe).

Combinaison avec les modèles les plus différents (II)

Jeu de données	$\tilde{\beta}_m$	modèle EDDA
bupa	0.4261	0.4000
crabes	0.0500	0.0500
haberman	0.2516	0.2516
haemo	0.1600	0.1467
iris	0.0400	0.0400
thyroïde	0.0977	0.0977
vin	0.0449	0.0449

Tableau 4.33 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *empilement* avec les modèles les plus diférents (II).

L'empilement avec la stratégie les modèles les plus différents (II) mérite la même critique que la statégie précédente (cf. tableau A.21 de l'annexe). Ainsi l'empilement s'avère dans le contexte étudié une technique de combinaison décevante.

4.4 Une incursion non paramétrique

Enfin, dans un souci de bien explorer les possibilités pratiques de la combinaison de modèles en analyse discriminante, nous sommes sortis du contexte exclusivement gaussien. Ainsi, nous avons associé par combinaison, une méthode non paramétrique, d'inspiration locale, la méthode des k-plus proches voisins (k-ppv) avec deux méthodes paramétriques, d'inspiration globale, l'analyse discriminante linéaire $[\lambda \mathbf{DAD}']$ et le classement à la moyenne la plus proche $[\lambda \mathbf{I}]$. Dans ce cadre nous avons essayé les techniques de combinaison,

- régression séparée pour chaque classe,
- régression vectorisée,
- comité de méthodes.

Les techniques qui reposent sur la vraisemblance (pure, pénalisée, intégrée ou empilement) ne sont pas praticables puisque la vraisemblance n'est pas disponible pour les modèles non paramétriques.

Nous présentons, d'abord, le TEC associé au meilleur modèle parmi les trois modèles en question : (k-ppv), $[\lambda \mathbf{DAD}']$ et $[\lambda \mathbf{I}]$ (tableau 4.34).

Jeu de données	Modèle	TEC
bupa	$[\lambda\mathbf{DAD}']$, $[k-\text{ppv}]$	0.3043
crabes	$[\lambda\mathbf{DAD}']$	0.0500
haberman	$[\lambda\mathbf{DAD}']$	0.2516
haemo	$[\lambda\mathbf{DAD}']$	0.1467
iris	$[\lambda\mathbf{DAD}']$, $[k-\text{ppv}]$	0.0200
thyroïde	$[k-\text{ppv}]$	0.0837
vin	$[\lambda\mathbf{DAD}']$	0.0112

Tableau 4.34 : Meilleur modèle d'analyse discriminante parmi les méthodes des (k-ppv), $[\lambda\mathbf{DAD}']$ et $[\lambda\mathbf{I}]$ et taux d'erreur de classement estimé pour chaque jeu de données.

Régression séparée

La combinaison par régression séparée avec cette stratégie fait aussi bien que le meilleur modèle présent dans la combinaison pour les jeux de données **crabes**, **haemo**, **iris**, **thyroïde** et **vin** (tableau 4.35). Le tableau 4.36 exhibe les modèles contribuant à la combinaison et les poids respectifs pour quelques jeux de données. Une fois de plus, pour **haemo**, seul le modèle d'analyse discriminante linéaire est retenu. Pour **iris** il est intéressant que le groupe 1 - *Setosa* - soit séparé par la méthode des k-ppv tandis que les autres groupes sont séparés par le modèle d'analyse discriminante linéaire.

Jeu de données	contrainte de non négativité					modèle choisi
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3246	0.3246	0.3246	0.3246	0.3246	0.3043
crabes	0.0600	0.0500	0.0500	0.0550	0.0550	0.0500
haberman	0.2484	0.2484	0.2484	0.2582	0.2582	0.2451
haemo	0.1600	0.1600	0.1600	0.1467	0.1467	0.1467
iris	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200
thyroïde	0.0884	0.0837	0.0837	0.0837	0.0930	0.0837
vin	0.0112	0.0112	0.0112	0.0112	0.0112	0.0112

Tableau 4.35 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression separée.*

Jeu de données	groupe régressé	modèles contribuant à la combinaison et poids respectifs	modèle choisi
haemo	groupe 1	$[\lambda \mathbf{DAD}']$	$[\lambda \mathbf{DAD}']$
	groupe 2	$[\lambda \mathbf{DAD}']$	
iris	groupe 1	$(k-\text{ppv})$	$[\lambda \mathbf{DAD}']$, $(k-\text{ppv})$
	groupe 2	$[\lambda \mathbf{DAD}']$	
	groupe 3	$[\lambda \mathbf{DAD}']$	
thyroïde	groupe 1	$(k-\text{ppv})$	$(k-\text{ppv})$
	groupe 2	$(k-\text{ppv})$	
	groupe 3	$0.18[\lambda \mathbf{DAD}'] + 0.22[\lambda \mathbf{I}] + 0.60(k-\text{ppv})$	

Tableau 4.36 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les méthodes $(k-\text{ppv})$, $[\lambda \mathbf{DAD}']$ et $[\lambda \mathbf{I}]$.

Régression vectorisée

Aussi pour cette stratégie et par la méthode régression vectorisée, ce sont **crabes**, **haemo**, **iris** et **thyroïde** qui produisent des TEC comparables au meilleur modèle parmi les modèles considérés par la combinaison.

Jeu de données	contrainte de non négativité					modèle choisi
	$\tilde{\beta}_m$	$\tilde{\beta}_m > 0.1$	$\tilde{\beta}_m > 0.1$ arrondis	$\tilde{\beta}_m > 0.1$ normalisés	contrainte unitaire	
bupa	0.3246	0.3246	0.3246	0.3246	0.4290	0.3043
crabes	0.0550	0.0500	0.0500	0.0550	0.0850	0.0500
haberman	0.2549	0.2647	0.2647	0.2451	0.4314	0.2451
haemo	0.1600	0.1467	0.1600	0.1467	0.2667	0.1467
iris	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200
thyroïde	0.0837	0.0837	0.0837	0.0837	0.0930	0.0837
vin	0.0112	0.0112	0.0112	0.0112	0.0112	0.0112

Tableau 4.37 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *régression vectorisée.*

Comité de méthodes

Par la technique de combinaison du comité de méthodes les résultats concernant **haemo**, **iris** et **thyroïde** sont du même ordre que par la régression vectorisée.

Jeu de données	$\tilde{\beta}_m$	modèle choisi
bupa	0.3246	0.3043
crabes	0.0550	0.0500
haberman	0.2582	0.2451
haemo	0.1467	0.1467
iris	0.0200	0.0200
thyroïde	0.0837	0.0837
vin	0.0112	0.0112

Tableau 4.38 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *comité de méthodes*.

4.5 Performance de la méthode de combinaison hiérarchique

La combinaison hiérarchique concerne les problèmes dont le nombre de groupes est supérieur à deux. Ainsi, de notre liste, seulement les jeux de données qui sont dans ces conditions interviennent. Le tableau 4.39 ci-dessous présente les TEC respectifs associés au modèle de combinaison hiérarchique.

Jeu de données	combinaison hiérarchique	EDDA
crabes	0.0450	0.0450
iris	0.0200	0.0200
thyroïde	0.0372	0.0326
vin	0	0

Tableau 4.39 : Taux d'erreur de classement pour chaque jeu de données pour la méthode de combinaison *hiérarchique*.

Par cette technique les TEC sont égales aux TEC des modèle EDDA pour tous les cas à l'exception de **thyroïde**. Mais surtout la combinaison hiérarchique est attractive par sa simplicité et son interprétabilité. Dans la suite nous présentons les modèles discriminants par combinaison hiérarchique pour **iris** (figure 4.2) et **crabes** (figure 4.3). Ces deux figures sont accompagnées, à droite ou en-dessous, par la représentation des données sur leurs premier et deuxième axes factoriels (reprise de la figure 4.1).

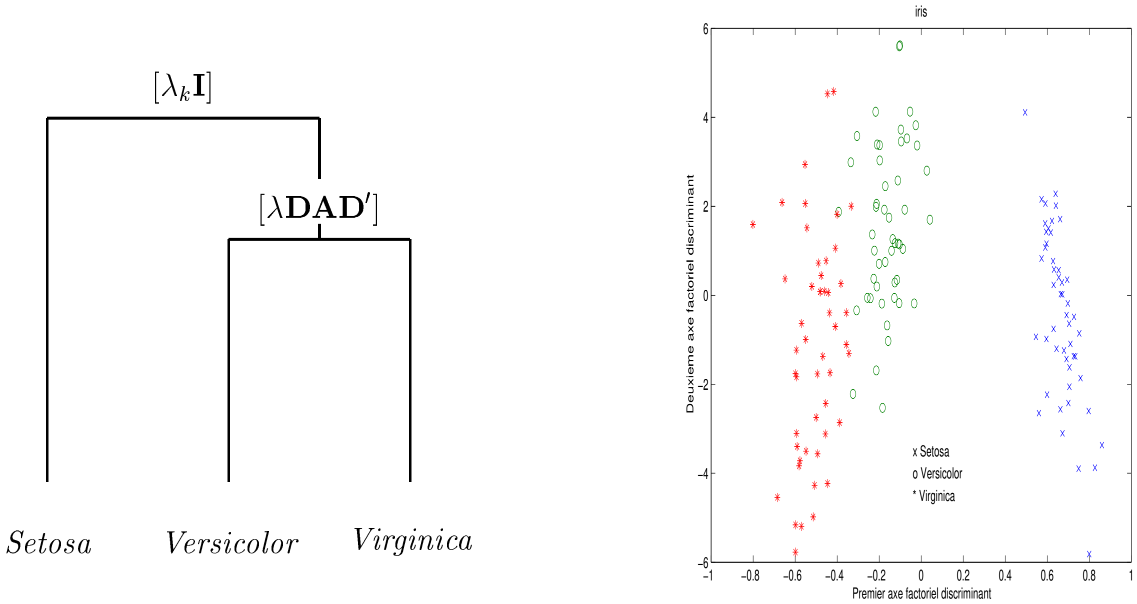


Figure 4.2 : Représentations du modèle de combinaison hiérarchique (à gauche) et des données sur les premier et deuxième axes factoriels discriminants (à droite) pour **iris** .

La lecture du modèle hiérarchique est la suivante. Au premier niveau de l'arbre c'est le modèle $[\lambda_k \mathbf{I}]$ - un modèle sphérique avec différents volumes - qui discrimine au mieux *Setosa* du groupe {*Versicolor, Virginica*}. Au deuxième niveau c'est le modèle linéaire $[\lambda \mathbf{DAD}']$ qui discrimine au mieux entre *Versicolor* et *Virginica.*

Notons que la procédure hiérarchique founit une règle de décision assez simple pour séparer *Setosa* et {*Versicolor, Virginica*}. En effet, ce modèle s'adapte bien à la structure des données tel qu'elle est représentée sur leurs axes factoriels ; une fois que *Setosa* est séparée des autres classes il suffit d'un simple modèle linéaire pour terminer la discrimination.

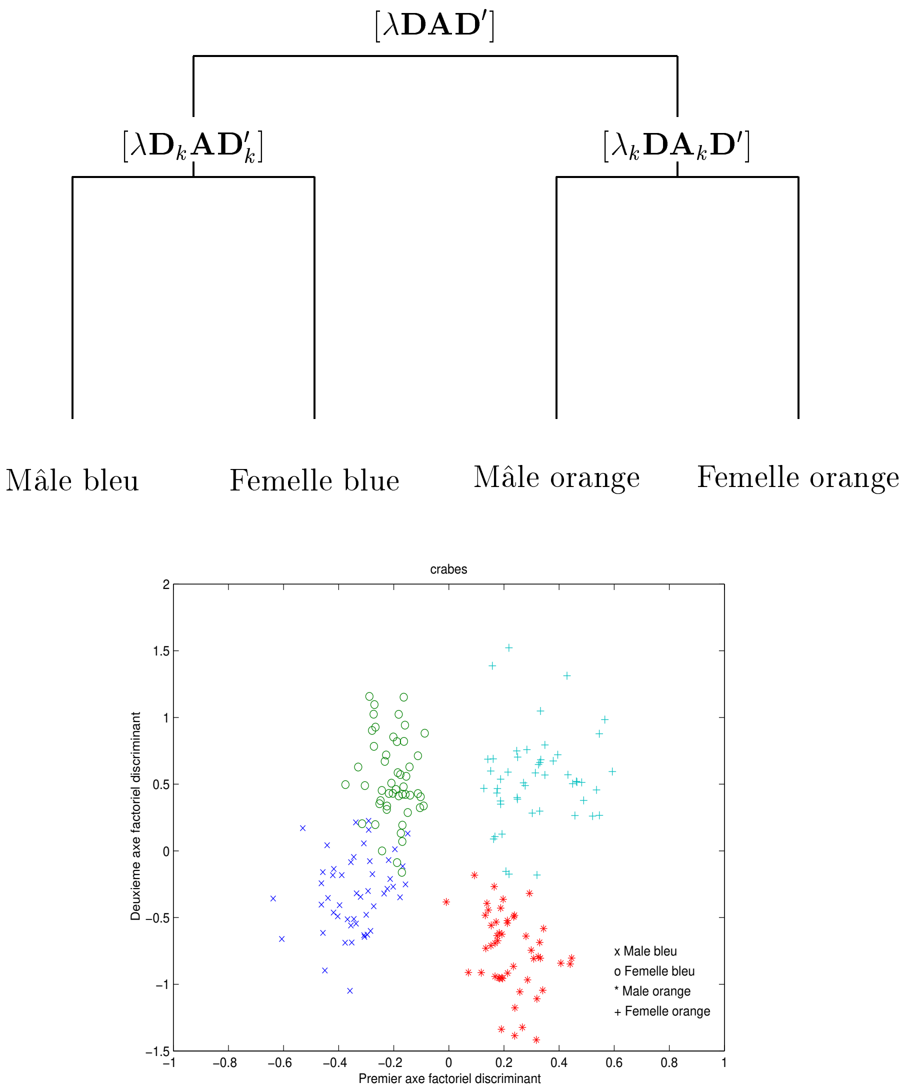


Figure 4.3 : Représentations du modèle de combinaison hiérarchique (en haut) et des données sur les premier et deuxième axes factoriels discriminants (en bas) pour **crabes**.

Pour les données **crabes**, la combinaison hiérarchique sépare, au premier niveau, les espèces et choisit le modèle linéaire. Au deuxième niveau le modèle $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ - un modèle elliptique avec différentes orientations - sépare les sexes dans les bleus. Au même niveau c'est le modèle $[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$ - un modèle elliptique avec la même orientation qui sépare les mâles et les femelles oranges. Une fois de plus la procédure hiérarchique est suggestive par rapport à la structure des données et donne une lecture intéressante de la discrimination.

4.6 Comparaison des méthodes de combinaison

Dans cette section nous proposons une comparaison des différentes techniques de combinaison par jeu de données. Pour chaque technique nous avons choisi la stratégie qui a été la plus performante. La stratégie englobant le modèle non paramétrique des k plus proches voisins a été prise en compte quand il s'agit des techniques de combinaison qui la concernent.

4.6.1 bupa

Pour le jeu de données **bupa** la régression séparée et le comité de méthodes produisent des TEC comparables au TEC associé au modèle d'analyse discriminante linéaire (tableau 4.40). Les deux techniques proposent une combinaison de trois modèles : le linéaire avec un poids, d'à peu près 40%, un modèle elliptique dont la pondération est sensiblement de 50% et finalement le modèle du classement à la moyenne la plus proche qui pèse 10%. Les technique des ratios des vraisemblances estimées et pénalisées avec une performance comparable à la stratégie EDDA ne proposent qu'un seul modèle, précisément l'analyse discriminante linéaire.

Technique discriminante	modèle utilisé		TEC
EDDA	$[\lambda\mathbf{DAD}']$		0.3043
Régression séparée			0.3043
groupe 1	$0.37[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}_k'] + 0.10[\lambda\mathbf{I}]$		
groupe 2	$0.39[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}_k'] + 0.13[\lambda\mathbf{I}]$		
Régression vectorisée			0.3159
	groupe 1	groupe 2	
groupe 1	$.40[\lambda\mathbf{DAD}']$		
	$+.60[\lambda\mathbf{D}_k\mathbf{AD}_k']$		
groupe 2		$.34[\lambda\mathbf{DAD}']$	
		$+.66[\lambda\mathbf{D}_k\mathbf{AD}_k']$	
Comité de méthodes	$0.41[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}_k'] + 0.11[\lambda\mathbf{I}]$		0.3043
Vraisemblance			
estimée	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$		0.3855
pénalisée	$[\lambda\mathbf{DAD}']$		0.3043
intégrée	$[\lambda\mathbf{DAD}']$		0.3043
Empilement	$0.42[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.48[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$		0.3855

Tableau 4.40 : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **bupa**.

4.6.2 crabes

Pour ce jeu de données la méthode de combinaison par les ratios de la vraisemblance pénalisée fait aussi bien que le meilleur modèle EDDA - un modèle elliptique avec différentes orientations - (tableau 4.41). Or il s'avère

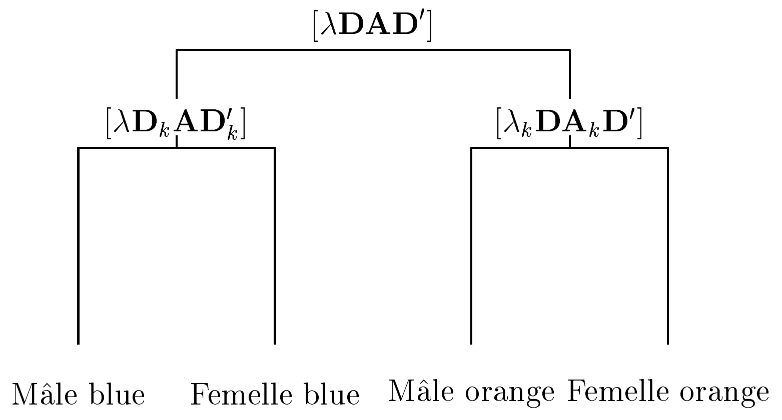
Technique discrimin.	modèle utilisé				TEC
EDDA	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k{}']$				0.0450
Régres. vectorisée	groupe 1	groupe 2	groupe 3	groupe 4	0.0500
groupe 1	$.59[\lambda \mathbf{DAD}']+$ $.41[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$				
groupe 2		$.75[\lambda \mathbf{DAD}']+$ $.25[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			
groupe 3			$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$		
groupe 4				$.5[\lambda \mathbf{DAD}']+$ $.5[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	
Hiérar.					0.0450

Tableau 4.41(a) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **crabes**.

que par la technique de combinaison mentionnée on est conduit au même modèle. Le modèle de combinaison hiérarchique a une bonne performance et propose une séparation assez intéressante des groupes (cf. section 4.5).

Technique discriminante	modèle utilisé	TEC
Régression séparée		0.0500
groupe 1	$0.57[\lambda\mathbf{DAD}'] + 0.43[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
groupe 2	$0.74[\lambda\mathbf{DAD}'] + 0.26[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
groupe 3	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
groupe 4	$0.45[\lambda\mathbf{DAD}'] + 0.55[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
Comité de méthodes	$0.16[\lambda_k\mathbf{DAD}'] + 0.83[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	0.0500
Vraisemblance		
estimée	$[\lambda_k\mathbf{B}]$	0.0500
pénalisée	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	0.0450
intégrée	$.65[\lambda\mathbf{D}_k\mathbf{AD}'_k] + .34[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	0.0500
Empilement	$[\lambda\mathbf{DAD}']$	0.0500

Tableau 4.41(b) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **crabes**.

4.6.3 haberman

Pour le jeu de données **haberman** aucune méthode de combinaison n'égale la stratégie EDDA en ce qui concerne la performance (tableau 4.42).

Technique discriminante	modèle utilisé		TEC
EDDA	$[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$		0.2418
Régression séparée			0.2516
groupe 1	$0.28[\lambda \mathbf{DAD}'] + 0.20[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + 0.52[\lambda \mathbf{I}]$		
groupe 2	$0.32[\lambda \mathbf{DAD}'] + 0.16[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + 0.52[\lambda \mathbf{I}]$		
Régression vectorisée			0.2451
	groupe 1	groupe 2	
groupe 1	$.51[\lambda \mathbf{DAD}']$ $+.28[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$+.21[\lambda \mathbf{I}]$	
groupe 2		$.57[\lambda \mathbf{DAD}']$ $+.43[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	
Comité de méthodes	$0.29[\lambda \mathbf{DAD}'] + 0.16[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + 0.54[\lambda \mathbf{I}]$		0.2516
Vraisemblance			
estimée	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k{}']$		0.2451
pénalisée	$.17[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}'] + .51[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + .31[\lambda_k \mathbf{B}_k]$		0.2516
intégrée	$.36[\lambda \mathbf{DAD}'] + .64[\lambda_k \mathbf{DAD}']$		0.2516
Empilement	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k{}']$		0.2451

Tableau 4.42 : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **haberman**.

4.6.4 haemo

Les méthodes de combinaison fondées sur la vraisemblance pénalisée et intégrée proposent deux modèles d'analyse discriminante dont les TEC associés sont comparables au TEC du modèle linéaire (tableau 4.43). Par la vraisemblance intégrée le modèle d'analyse discriminante est composé en 51%

Technique discriminante	modèle utilisé		TEC
EDDA	$[\lambda\mathbf{DAD}']$		0.1467
Régression séparée			0.1600
groupe 1	$[\lambda\mathbf{DAD}']$		
groupe 2	$[\lambda\mathbf{DAD}']$		
Régression vectorisée			0.1600
	groupe 1	groupe 2	
groupe 1	$[\lambda\mathbf{DAD}']$		
groupe 2		$[\lambda\mathbf{DAD}']$	
Comité de méthodes	$[\lambda\mathbf{DAD}']$		0.1600
Vraisemblance			
estimée	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$		0.1600
pénalisée	$.14[\lambda\mathbf{DAD}'] + .3[\lambda_k\mathbf{DAD}'] + .13[\lambda\mathbf{DA}_k\mathbf{D}'] +0.3[\lambda_k\mathbf{DA}_k\mathbf{D}'] + .13[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$		0.1467
intégrée	$.51[\lambda\mathbf{DAD}'] + .34[\lambda_k\mathbf{DAD}'] + 0.15[\lambda\mathbf{DA}_k\mathbf{D}']$		0.1467
Empilement	$[\lambda\mathbf{DAD}']$		0.1600

Tableau 4.43 : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **haemo**.

du modèle linéaire, en 34% d'un modèle dont les volumes sont différents et 15% d'un modèle de formes différentes.

4.6.5 iris

Seuls les méthodes de combinaison établies à partir de la vraisemblance soit estimée, pénalisée ou intégrée conduissent à des TEC plus grandes que le modèle linéaire (tableau 4.44). Tous les autres méthodes de combinaison lui sont comparable. La combinaison hiérarchique propose un modèle assez intéressant qui a été détaillé en section 4.5

.

Technique discriminante	modèle utilisé	TEC
Régression séparée		0.0200
groupe 1	$[\lambda_k \mathbf{B}]$	
groupe 2	$0.3[\lambda_k \mathbf{D}\mathbf{A}_k\mathbf{D}'] + 0.7[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k']$	
groupe 3	$0.24[\lambda_k \mathbf{D}\mathbf{A}_k\mathbf{D}'] + 0.76[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k']$	
Comité de méthodes	$0.32[\lambda_k \mathbf{D}\mathbf{A}_k\mathbf{D}'] + 0.68[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k']$	0.0200
Vraisemblance		
estimée	$[\lambda_k \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$	0.0267
pénalisée	$.58[\lambda_k \mathbf{D}_k\mathbf{A}\mathbf{D}_k{}'] + .42[\lambda_k \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k']$	0.0267
intégrée	$[\lambda_k \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k{}']$	0.0267
Empilement	$0.32[\lambda_k \mathbf{D}\mathbf{A}_k\mathbf{D}'] + 0.68[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k']$	0.0200

Tableau 4.44(a) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **iris**.

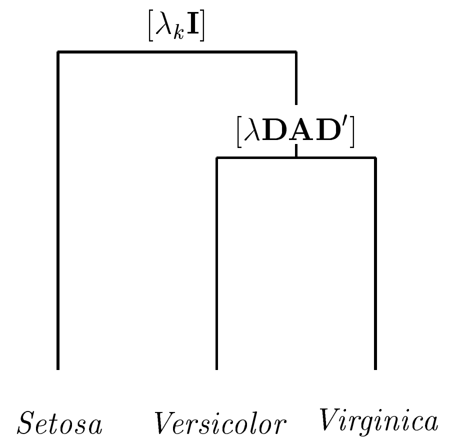
Technique discriminante	modèle utilisé			TEC
EDDA	$[\lambda\mathbf{DAD}']$			0.0200
Régression vectorisée				0.0200
	groupe 1	groupe 2	groupe 3	
groupe 1	$[\lambda\mathbf{B}]$			
groupe 2		$.3[\lambda_k\mathbf{DA}_k\mathbf{D}']$ $+.7[\lambda\mathbf{D}_k\mathbf{AD}'_k]$		
groupe 3			$.24[\lambda_k\mathbf{DA}_k\mathbf{D}']$ $+.76[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
Hiérarchique				0.0200

Tableau 4.44(b) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **iris**.

4.6.6 thyroïde

Pour ce jeu de données seul la méthode de combinaison fondée sur la vraisemblance intégrée a un TEC associé égal au meilleur modèle EDDA (tableau 4.45). Pourtant on vérifie que cette méthode de combinaison ne propose qu'un seul modèle et il est justement le meilleur de EDDA - un modèle diagonal avec différents volumes.

Technique discriminante	modèle utilisé	TEC
EDDA	$[\lambda_k \mathbf{B}]$	0.0326
Régression séparée		0.0372
groupe 1	$0.15[\lambda \mathbf{DAD}'] + 0.85[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	
groupe 2	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	
groupe 3	$0.23[\lambda \mathbf{DAD}'] + 0.65[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k'] + 0.12[\lambda \mathbf{I}]$	
Comité de méthodes	$0.30[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k'] + 0.68[\lambda_k \mathbf{I}]$	0.0372
Vraisemblance		
estimée	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0372
pénalisée	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0372
intégrée	$[\lambda_k \mathbf{B}]$	0.0326
Empilement	$0.30[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k'] + 0.68[\lambda_k \mathbf{I}]$	0.0372

Tableau 4.45(a) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **thyroïde**.

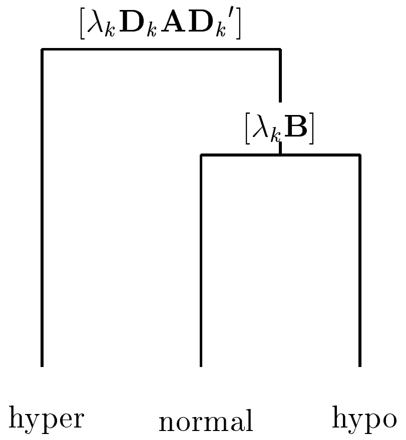
Technique discriminante	modèle utilisé			TEC
Régression vectorisée				0.0372
	groupe 1	groupe 2	groupe 3	
groupe 1	$.14[\lambda \mathbf{DAD}']$ $+.86[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			
groupe 2		$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$		
groupe 3			$.28[\lambda \mathbf{DAD}']$ $+.72[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	
Hiérarchique				0.0372

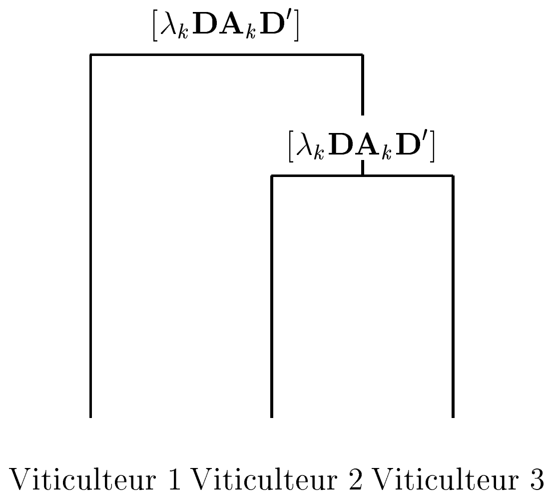
Tableau 4.45(b) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **thyroïde**.

4.6.7 vin

Pour le jeu de données **vin** le meilleur modèle EDDA - un modèle avec différents volumes et différentes orientations - a un TEC nul. En combinaison c'est la méthode hiérarchique qui conduit au même TEC. Celle-ci sépare d'abord le groupe concernant le viticulteur 1 des autres groupes par un modèle elliptique avec différents volumes et différentes formes. Ensuite le même modèle

Technique discriminante	modèle utilisé	TEC
EDDA	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k{}']$	0
Régression séparée		0.0056
groupe 1	$0.19[\lambda \mathbf{DAD}'] + 0.35[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$	
groupe 2	$0.18[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k'] + 0.27[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$	
groupe 3	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$	
Comité de méthodes	$0.1[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}'] + 0.25[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$ $+0.65[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$	0.0056
Vraisemblance		
estimée	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0056
pénalisée	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	0.0056
intégrée	$[\lambda \mathbf{DAD}']$	0.0112
Empilement	$0.27[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}'] + .29[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k{}']$ $+0.22[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k{}']$	0.0056

Tableau 4.46(a) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **vin**.

<table>
<tr><th>Technique discriminante</th><th>modèle utilisé</th><th>TEC</th></tr>
<tr><td>Régression vectorisée</td><td></td><td>0.0056</td></tr>
<tr><td>groupe 1</td><td>$.19[\lambda \mathbf{DAD}']$
$+.35[\lambda_k \mathbf{DA}_k \mathbf{D}']$
$+.18[\lambda \mathbf{D}_k \mathbf{AD}_k{}']$
$+.27[\lambda_k \mathbf{D}_k \mathbf{AD}_k{}']$</td><td></td></tr>
<tr><td>groupe 2</td><td>$.28[\lambda_k \mathbf{DA}_k \mathbf{D}']$
$+.72[\lambda_k \mathbf{D}_k \mathbf{AD}'_k]$</td><td></td></tr>
<tr><td>groupe 3</td><td>$[\lambda_k \mathbf{D}_k \mathbf{AD}'_k]$</td><td></td></tr>
<tr><td>Hiérarchique</td><td>


</td><td>0</td></tr>
</table>

Tableau 4.46(b) : Modèles contribuant à la combinaison et poids respectifs et taux d'erreur de classement associé par les différentes techniques de combinaison pour le jeu de données **vin**.

est utilisé pour séparer les groupes des viticulteurs 2 et 3. Encore une fois il s'agit d'une lecture suggestive alliée à une bonne performance pour la combinaison hiérarchique.

4.7 Reconnaissance de tumeurs cérébrales

Nous menons une application sur des données aimablement fournis par l'Hôpital Universitaire de Grenoble[2]. L'objectif est la caractérisation tissulaire des tumeurs cérébrales *in vivo* chez l'homme. Le diagnostic repose sur des variables mesurées par imagerie spectroscopique, outil particulier, de la spectroscopie de résonance magnétique nucléaire du proton.

Comme Szabo, Esteve, Grand, Décops, Lefournier, Le Bas, and Rémy (2000) nous l'explique, la caractérisation histopathologique des tumeurs du cerveau est faite après la réalisation d'une biopsie. Cette procédure est ce que l'on appelle en médecine une procédure du type invasive. Puisque cette procédure peut entraîner des conséquences néfastes, des techniques de diagnostic des tumeurs cérébrales du type non invasive ont émergé. Dans ce cadre l'imagerie par résonance magnétique spectroscopique (IRMS) a prouvé son utilité. La question qui reste ouverte est comment utiliser cette information de façon à mieux diagnostiquer les différentes sortes de tumeurs.

La spectroscopie de résonance magnétique est une technique qui permet d'obtenir pour un voxel - un volume élémentaire de tissu - un spectre caractéristique. Un spectre comporte des pics correspondant chacun à une ou plusieurs substances chimiques particulières identifiables par leur position en fréquence. On identifie habituellement, dans les spectres proton du cerveau et de tumeurs *in vivo* chez l'homme, 6 pics principaux. Leur description et leur signification biologique associée sont bien expliqués par Szabo (2001).

[2]Unité mixte INSERM - Université Joseph Fourrier, U438 , Centre Hospitalier Universitaire de Grenoble

Au début la spectroscopie mono-voxel était utilisée. La classification reposait sur l'étude des ratios des pics. Ensuite, tout le spectre après digitalisation, ou un profil des intensités de pics sélectionnés ont été utilisés. Ces méthodes ont eu beaucoup de succès. Pourtant la distinction entre deux types de tumeurs : les gliomes de haut-grade et les gliomes de bas-grade restait douteuse. Du fait de l'acquisition d'un seul voxel, ces techniques ne prennent pas en compte l'hétérogénéité du tissu cérébral. Ainsi une autre approche pour grouper les tumeurs cérébrales est apparue. Pour celle-ci les spectres de tous les voxels d'une région d'intérêt sélectionnée sont considéreés simultanément. Mais la quantité d'information obtenue est immense et difficile à analyser.

Une autre approche pour tirer l'information des données de l'IRMS a été développée par Szabo, Esteve, Grand, Décops, Lefournier, Le Bas, and Rémy (2000). Chaque voxel d'une image spectroscopique est attribué à un groupe. Ces données sont ensuite résumées sur une seule image dont chaque pixel est colorié selon le groupe qui lui est attribuée. L'objectif est de faire de cette image un instrument de diagnostic facile à manipuler par les radiologistes.

Le jeu de données, qu'on notera **cerveau**, est composé de 77 observations mesurées sur 7 variables et réparties en 7 groupes. Les groupes sont les 4 types histopathologiques de tumeurs, définis par l'Organisation Mondiale de la Santé,

- **BG** gliomes de bas-grade,
- **HG** gliomes de haut-grade,
- **Méta** métastases,
- **ME** méningiomes,

et 3 autres types de tissus cérébraux,

- **NE** nécrose,
- **LCR** liquide céphalo-rachidien,
- **NT** tissu cérébral non tumoral.

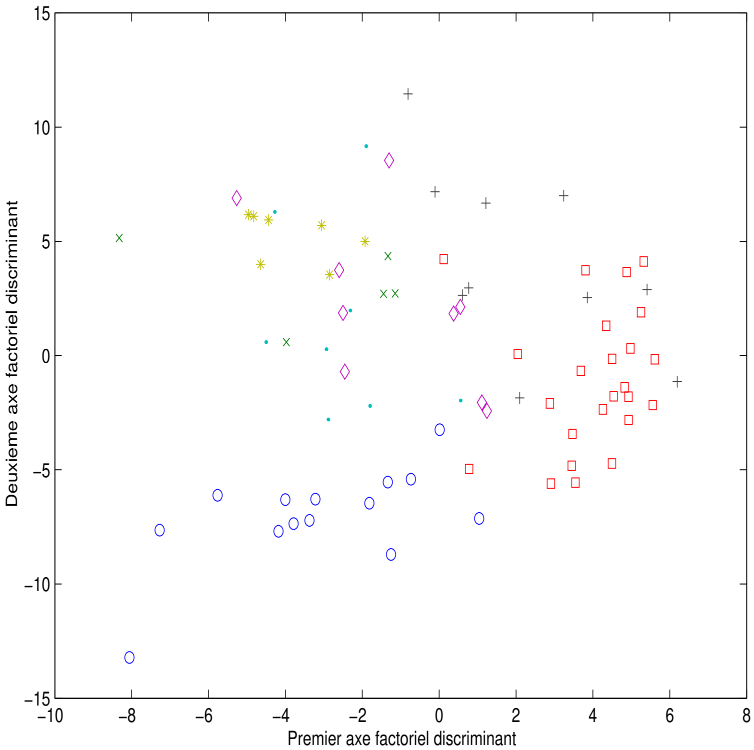
Les variables qui caracterisent les groupes sont les aires des six plus grands pics du spectre moyen d'une région d'intérêt séléctionnée localisée dans la tumeur. Un autre variable, ne provenant pas des données spectroscopiques, est relative à l'intensité des images pondérées en T2 (cf. Szabo 2001).

Dans ce jeu de données la distribution des observations par groupes est la suivante,

BG	**HG**	**Méta**	**ME**	**NE**	**LCR**	**NT**
14	24	10	8	9	5	7

Pour avoir une idée de la structure des groupes nous avons représenté les données sur les premier et deuxième axes factoriels discriminants (figure 4.4).

Nous avons appliqué l'analyse discriminante par la méthode de combinaison hiérarchique. Le modèle qui en résulte est représenté sur la figure 4.5. On remarque que le modèle sépare au premier et deuxième niveaux des groupes non tumoraux : **NT** et **LCR** ce qui est intéressant. De plus, ce modèle remet la séparation entre **HG** et **Méta** au dernier niveau. En effet il s'avère que la distinction entre ces deux groupes est difficile en termes de spectre (cf. Szabo 2001).



BG	HG	Méta	ME	NE	LCR	NT
o	□	▽	·	◇	x	*

Figure 4.4 : Représentation de **cerveau** sur les premier et deuxième axes factoriels discriminants.

De la lecture de la figure 4.5 on peut également remarquer que les méthodes choisis par la combinaison hiérarchique sont très simples. La méthode la plus complexe séléctionnée est l'analyse discriminante linéaire. Les autres méthodes supposent des structures diagonales des matrices de variance. Cela conforte dans l'idée que les groupes sont assez faciles à reconnaître et que les erreurs de classement sont pour la plupart dues à des erreurs d'étiquetage de l'échantillon d'apprentissage à cause de la difficulté de discerner les spectres.

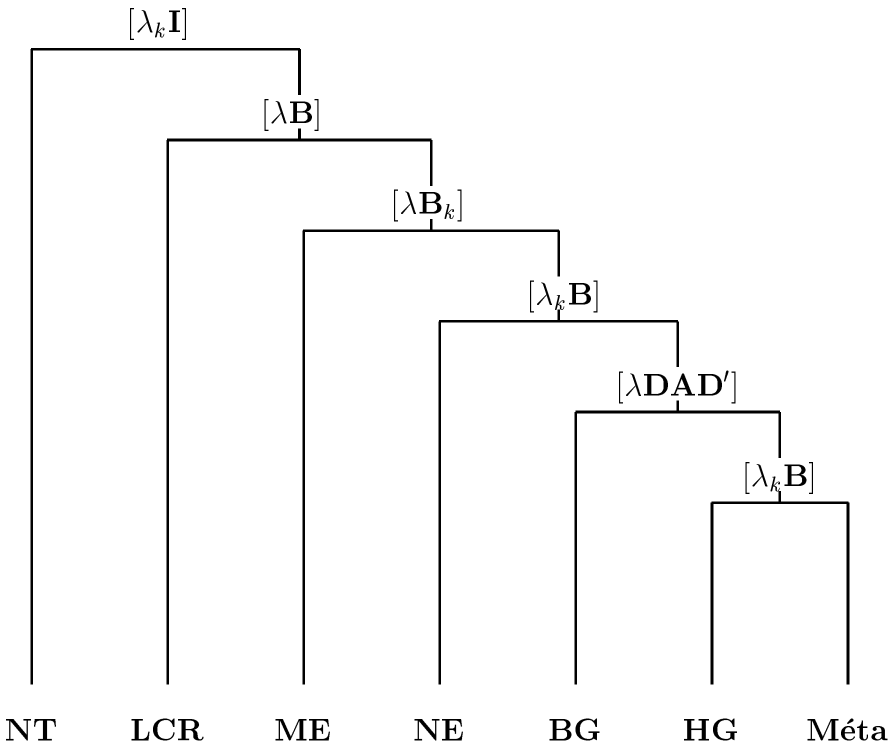


Figure 4.5 : Représentation du modèle de combinaison hiérarchique pour le jeu de données **cerveau**.

Par méthode de combinaison hiérarchique le taux d'erreur de classement (TEC) obtenu par validation croisée est de 0.1429. Le tableau 4.47 montre la matrice de confusion qui résulte de la méthode de combinaison hiérarchique.

Groupe d'origine		groupe d'affectation							
		BG	**HG**	**Méta**	**ME**	**NE**	**LCR**	**NT**	**HG,Méta**
BG	14	13							1
HG	24		19	3	1	1			
Méta	10		4	5		1			
ME	8				8				
NE	9					9			
LCR	5						5		
NT	7							7	

Tableau 4.47 : Matrice de confusion pour **cerveau** par la méthode de combinaison hiérarchique.

L'analyse discriminante linéaire utilisée par Szabo (2001) produit un TEC, aussi par validation standard, de 0.1299. La matrice de confusion correspondante est exhibée sur le tableau 4.48.

Groupe d'origine		groupe d'affectation						
		BG	**HG**	**Méta**	**ME**	**NE**	**LCR**	**NT**
BG	14	13	1					
HG	24	1	19	3		1		
Méta	10		3	6		1		
ME	8				8			
NE	9					9		
LCR	5						5	
NT	7							7

Tableau 4.48 : Matrice de confusion pour **cerveau** par la méthode d'analyse discriminante linéaire.

Si, comme le fait Szabo (2001) nous regroupons **HG** et **Méta** en en seul groupe représentant les tumeurs agressives **Agr** on obtient la matrice de confusion suivante,

		BG	**Agr**	**ME**	**NE**	**LCR**	**NT**
BG	14	13	1				
Agr	34		31	1	2		
ME	8			8			
NE	9				9		
LCR	5					5	
NT	7						7

Tableau 4.49 : Matrice de confusion pour **cerveau** par la méthode de combinaison hiérarchique en regroupant **HG** et **Méta**.

De cette façon la méthode de combinaison hiérarchique conduit au même taux d'erreur que l'analyse discriminante linéaire. La différence entre les deux méthodes se porte sur le classement attribué à un élément de **Agr**. Notre méthode dit qu'il s'agit d'un méningiome (**ME**) tandis que le modèle linéaire le confond avec une tumeur de bas-grade (**BG**).

Remarque : pour ce jeu de données aucune analyse n'a pris en compte la considération fort plausible de côuts de mauvais classement différents selon les groupes.

Conclusions et perspectives

Nous terminons cette thèse par un bilan résumant les conclusions que nous tirons de cette étude et dressons quelques perspectives.

Nous avons analysé sur des jeux de données publiés sept techniques de combinaison en série et une technique de combinaison hiérarchique.

Pour le premier type de combinaison les modèles sont combinés de façon convexe en général sous une forme linéaire. Trois de ces méthodes, la régression séparée par groupe, la régression par vectorisation de la matrice des densités et le comité de méthodes reposent sur le critère des moindres carrés. Leur but commun est la minimisation de l'erreur de classement. Ces techniques se distinguent entre elles par l'expression du critère d'erreur à minimiser.

Les trois autres méthodes s'appuient sur la vraisemblance des différents modèles. Ce sont les ratios des vraisemblances, les ratios des vraisemblances pondérées, les ratios des vraisemblances intégrées et l'empilement. Les ratios sont des rapports entre les vraisemblances mentionnées tandis que l'empilement utilise dans sa procédure l'algorithme EM.

La régression séparée par groupe et la régression par vectorisation estiment trop de coefficients. En effet, pour ces techniques il y a un coefficient par modèle inclus dans la combinaison et ce pour chaque groupe. Son interprétation devient compliquée et devient même impossible.

Toutes les autres techniques de combinaison en série analysées estiment un seul coefficient par modèle. Les ratios des vraisemblances ont le désavantage de pondérer énormément le modèle le plus complexe. Cet inconvénient est surmonté par l'utilisation de la vraisemblance pondérée. Cette méthode, ainsi que les ratios des vraisemblances intégrées proposent souvent un modèle unique. Ce comportement montre qu'il n'y a pas lieu d'attendre d'amélioration des taux de reconnaissance par une stratégie de combinaison pour les jeux de données que nous avons considéré. C'est sans doute un atout de ces techniques.

En général les taux d'erreur de classement associées aux méthodes de combinaison en série sont assez décevants. En effet, ils ne sont pas meilleurs que ceux du modèle le plus performant. Il faut pourtant remarquer que nous avons évalué l'espérance de l'estimateur du taux d'erreur de classement. Il reste à évaluer la variance de cet estimateur. Nous sommes persuadés que la combinaison de modèles permettra réduire cette variance. Et cela à cause de l'effet de "moyenne" provoqué par la combinaison. Si cette conjecture qui reste à prouver par le moyen de simulations se vérifie, alors le fait d'obtenir des espérances du taux d'erreur semblables à celles du meilleur modèle n'est pas aussi décevant. L'important est que la combinaison de modèles fournit une règle de décision plus stable que chacune d'entre elles.

La combinaison hiérarchique s'est révélée en revanche particulièrement intéressante. Dans le cas où le nombre de groupes est supérieur à 2 cette technique de combinaison semble prometteuse par sa simplicité et sa facilité d'interprétation. Ces résultats confirment ceux obtenus par Ferreira (2001) pour la combinaison hiérarchique dans le contexte de l'analyse discriminante sur variables qualitatives. C'est le résultat le plus positif de notre recherche. Le couplage hiérarchique peut souvent conduire à des règles de décisions originales, parcimonieuses et très performantes. On peut donc recommander l'emploi de ce type d'approche pour des problèmes multi-groupes. De ce point de vue, une perspective intéressante serait d'explorer tous les arbres possibles pour encore améliorer les performances. Cela se fait aisément lorsque le nombre de groupes K est inférieur à cinq. Autrement, on peut envisager d'implanter un algorithme "branch and bound" pour garder cette recherche de l'arbre optimal faisable en un temps raisonnable.

Le travail que nous avons poursuivi nous a mené dans certains cas à des échecs et des déceptions, mais aussi à quelques résultats intéressants. Notre plus grande joie est de penser que nous avons quelque peu contribué à élargir les horizons de la recherche, dans le domaine de l'analyse discriminante.

Annexe

Modèles contribuant à la combinaison et poids respectifs

Jeu de données	groupe régressé	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	groupe 1	$0.39[\lambda\mathbf{DAD}'] + 0.5[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.11[\lambda\mathbf{I}]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.39[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.13[\lambda\mathbf{I}]$	
crabes	groupe 1	$0.17[\lambda\mathbf{DAD}'] + 0.14[\lambda_k\mathbf{DAD}']+$ $+0.3[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.39[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$
	groupe 2	$0.35[\lambda_k\mathbf{DAD}'] + 0.65[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 3	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 4	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
haemo	groupe 1	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
	groupe 2	$[\lambda\mathbf{DAD}']$	
haberman	groupe 1	$0.33[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.35[\lambda\mathbf{B}] + 0.32[\lambda\mathbf{I}]$	$[\lambda\mathbf{DA}_k\mathbf{D}']$
	groupe 2	$0.32[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.34[\lambda\mathbf{B}] + 0.34[\lambda\mathbf{I}]$	
iris	groupe 1	$[\lambda_k\mathbf{B}]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.3[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.7[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 3	$0.24[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.76[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	
thyroïde	groupe 1	$0.32[\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.68[\lambda_k\mathbf{B}_k]$	$[\lambda_k\mathbf{B}]$
	groupe 2	$[\lambda_k\mathbf{B}_k]$	
	groupe 3	$0.36[\lambda\mathbf{B}_k] + 0.64[\lambda_k\mathbf{B}_k]$	
vin	groupe 1	$0.19[\lambda\mathbf{DAD}'] + 0.35[\lambda_k\mathbf{DA}_k\mathbf{D}']$ $0.18[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.27[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}_k{}']$
	groupe 2	$0.28[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.72[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	
	groupe 3	$[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	

Tableau A.1 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les 14 modèles EDDA.

Jeu de données	groupe régressé	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	groupe 1	$0.72[\lambda\mathbf{DAD}'] + 0.28[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.73[\lambda\mathbf{DAD}'] + 0.27[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
crabes	groupe 1	$0.57[\lambda\mathbf{DAD}'] + 0.43[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.74[\lambda\mathbf{DAD}'] + 0.26[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
	groupe 3	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
	groupe 4	$0.45[\lambda\mathbf{DAD}'] + 0.55[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
haberman	groupe 1	$0.28[\lambda\mathbf{DAD}'] + 0.20[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.52[\lambda\mathbf{I}]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
	groupe 2	$0.32[\lambda\mathbf{DAD}'] + 0.16[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.52[\lambda\mathbf{I}]$	
haemo	groupe 1	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
	groupe 2	$[\lambda\mathbf{DAD}']$	
iris	groupe 1	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
	groupe 2	$0.51[\lambda\mathbf{DAD}'] + 0.39[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.1[\lambda\mathbf{I}]$	
	groupe 3	$0.59[\lambda\mathbf{DAD}'] + 0.41[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
thyroïde	groupe 1	$0.15[\lambda\mathbf{DAD}'] + 0.85[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
	groupe 2	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
	groupe 3	$0.23[\lambda\mathbf{DAD}'] + 0.65[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.12[\lambda\mathbf{I}]$	
vin	groupe 1	$0.57[\lambda\mathbf{DAD}'] + 0.43[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k{}']$
	groupe 2	$0.31[\lambda\mathbf{DAD}'] + 0.69[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
	groupe 3	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	

Tableau A.2 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les modèles les plus différents (I).

Jeu de données	groupe régressé	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	groupe 1	$0.76[\lambda\mathbf{B}] + 0.24[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}]$
	groupe 2	$0.78[\lambda\mathbf{B}] + 0.22[\lambda\mathbf{I}]$	
crabes	groupe 1	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{I}]$
	groupe 2	$[\lambda\mathbf{DAD}']$	
	groupe 3	$[\lambda\mathbf{DAD}']$	
	groupe 4	$[\lambda\mathbf{DAD}']$	
haberman	groupe 1	$0.43[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.57[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}_k]$
	groupe 2	$0.40[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.60[\lambda\mathbf{I}]$	
haemo	groupe 1	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
	groupe 2	$[\lambda\mathbf{DAD}']$	
iris	groupe 1	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
	groupe 2	$0.85[\lambda\mathbf{B}] + 0.15[\lambda\mathbf{I}]$	
	groupe 3	$0.82[\lambda\mathbf{B}] + 0.18[\lambda\mathbf{I}]$	
thyroïde	groupe 1	$0.79[\lambda\mathbf{B}] + 0.21[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}]$
	groupe 2	$0.76[\lambda\mathbf{B}] + 0.24[\lambda\mathbf{I}]$	
	groupe 3	$0.27[\lambda\mathbf{B}] + 0.73[\lambda\mathbf{I}]$	
vin	groupe 1	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
	groupe 2	$0.85[\lambda\mathbf{B}] + 0.15[\lambda\mathbf{I}]$	
	groupe 3	$0.82[\lambda\mathbf{B}] + 0.18[\lambda\mathbf{I}]$	

Tableau A.3 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les modèles les plus diférents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs				modèle EDDA
bupa groupe régressé	groupe 1		groupe 2		
groupe 1	$.40[\lambda\mathbf{DAD}']$ $+.60[\lambda\mathbf{D}_k\mathbf{AD}'_k]$				$[\lambda\mathbf{DAD}']$
groupe 2			$.34[\lambda\mathbf{DAD}']$ $+.66[\lambda\mathbf{D}_k\mathbf{AD}'_k]$		
crabes groupe régressé	groupe 1	groupe 2	groupe 3	groupe 4	
groupe 1	$.33[\lambda\mathbf{DAD}']$ $+.28[\lambda_k\mathbf{DA}_k\mathbf{D}']$ $+.38[\lambda\mathbf{D}_k\mathbf{AD}'_k]$				$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$
groupe 2		$.35[\lambda_k\mathbf{DAD}']$ $+.65[\lambda_k\mathbf{DA}_k\mathbf{D}']$			
groupe 3			$[\lambda_k\mathbf{DA}_k\mathbf{D}']$		
groupe 4				$[\lambda_k\mathbf{DA}_k\mathbf{D}']$	

Tableau A.4(a) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs		modèle EDDA
haberman			$[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$
groupe régressé	groupe 1	groupe 2	
groupe 1	$.38[\lambda \mathbf{DAD}']$ $+.20[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$ $+.20[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	$+.22[\lambda_k \mathbf{I}]$	
groupe 2		$.46[\lambda \mathbf{DAD}']$ $+.32[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k']$ $+.22[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	
haemo			$[\lambda \mathbf{DAD}']$
groupe régressé	groupe 1	groupe 2	
groupe 1	$[\lambda \mathbf{DAD}']$		
groupe 2		$[\lambda \mathbf{DAD}']$	

Tableau A.4(b) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs			modèle EDDA
iris				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$[\lambda \mathbf{B}]$			
groupe 2		$.3[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$ $+.7[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$		$[\lambda \mathbf{D}\mathbf{A}\mathbf{D}']$
groupe 3			$.24[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$ $+.76[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$	
thyroïde				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$.3[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$ $+.7[\lambda_k \mathbf{B}_k]$			
groupe 2		$[\lambda_k \mathbf{B}_k]$		$[\lambda_k \mathbf{B}_k]$
groupe 3			$.41[\lambda \mathbf{B}_k]$ $+.59[\lambda_k \mathbf{B}_k]$	
vin				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$.19[\lambda \mathbf{D}\mathbf{A}\mathbf{D}']$ $+.35[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$ $+.18[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}_k{}']$ $+.27[\lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}_k{}']$			
groupe 2		$.28[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$ $+.72[\lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$		$[\lambda \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$
groupe 3			$[\lambda_k \mathbf{D}_k \mathbf{A}\mathbf{D}'_k]$	

Tableau A.4(c) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant a la combinaison et poids respectifs				modèle EDDA
bupa groupe régres.	groupe 1	groupe 2			
groupe 1	$.72[\lambda \mathbf{DAD}']+$ $.28[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$				$[\lambda \mathbf{DAD}']$
groupe 2		$.73[\lambda \mathbf{DAD}']+$ $.27[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$			
crabes groupe régres.	groupe 1	groupe 2	groupe 3	groupe 4	
groupe 1	$.59[\lambda \mathbf{DAD}']+$ $.41[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$				
groupe 2		$.75[\lambda \mathbf{DAD}']+$ $.25[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$			$[\lambda \mathbf{DAD}']$
groupe 3			$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$		
groupe 4				$.5[\lambda \mathbf{DAD}']+$ $.5[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k']$	

Tableau A.5(a) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs		modèle EDDA
haberman			
groupe régressé	groupe 1	groupe 2	
groupe 1	$.51[\lambda\mathbf{DAD}']$ $+.28[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$+.21[\lambda\mathbf{I}]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
groupe 2		$.57[\lambda\mathbf{DAD}']$ $+.43[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	
haemo			
groupe régressé	groupe 1	groupe 2	
groupe 1	$[\lambda\mathbf{DAD}']$		$[\lambda\mathbf{DAD}']$
groupe 2		$[\lambda\mathbf{DAD}']$	

Tableau A.5(b) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs			modèle EDDA
iris				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			
groupe 2		$.51[\lambda \mathbf{DAD}']+$ $.39[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$ $+.1[\lambda \mathbf{I}]$		$[\lambda \mathbf{DAD}']$
groupe 3			$.59[\lambda \mathbf{DAD}']+$ $.4[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	
thyroïde				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$.14[\lambda \mathbf{DAD}']+$ $.86[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			
groupe 2		$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$		$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
groupe 3			$.28[\lambda \mathbf{DAD}']+$ $.72[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	
vin				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$.57[\lambda \mathbf{DAD}']+$ $.43[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$			
groupe 2		$.31[\lambda \mathbf{DAD}']+$ $.69[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$		$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
groupe 3			$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	

Tableau A.5(c) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs				modèle EDDA
bupa					
groupe régressé	groupe 1		groupe 2		
groupe 1	.48[$\lambda\mathbf{B}$] +.41[$\lambda\mathbf{I}$]		+.11[$\lambda_k\mathbf{I}$]		[$\lambda\mathbf{B}$]
groupe 2	+.12[$\lambda_k\mathbf{I}$]		.45[$\lambda\mathbf{B}$] +.43[$\lambda\mathbf{I}$]		
crabes					
groupe régressé	groupe 1	groupe 2	groupe 3	groupe 4	
groupe 1	[$\lambda\mathbf{DAD}'$]				
groupe 2		[$\lambda\mathbf{DAD}'$]			[$\lambda\mathbf{I}$]
groupe 3			[$\lambda_k\mathbf{DAD}'$]		
groupe 4				[$\lambda\mathbf{DAD}'$]	

Tableau A.6(a) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs		modèle EDDA
haberman			
groupe régressé	groupe 1	groupe 2	
groupe 1	$.42[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ $+.27[\lambda \mathbf{B}_k]$ $+.14[\lambda \mathbf{I}]$	$+.16[\lambda \mathbf{B}_k]$	$[\lambda \mathbf{B}_k]$
groupe 2		$.57[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ $+.24[\lambda \mathbf{B}_k]$ $+.19[\lambda \mathbf{I}]$	
haemo			
groupe régressé	groupe 1	groupe 2	
groupe 1	$[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$		$[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$
groupe 2		$[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$	

Tableau A.6(b) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs			modèle EDDA
iris				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$[\lambda \mathbf{B}]$			
groupe 2		$.85[\lambda \mathbf{B}]$ $+.15[\lambda \mathbf{I}]$		$[\lambda \mathbf{B}]$
groupe 3			$.81[\lambda \mathbf{B}]$ $+.19[\lambda \mathbf{I}]$	
thyroïde				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$.79[\lambda \mathbf{B}]$ $+.21[\lambda \mathbf{I}]$			
groupe 2		$.80[\lambda \mathbf{B}] + .20[\lambda \mathbf{I}]$		$[\lambda \mathbf{B}]$
groupe 3			$.34[\lambda \mathbf{B}]$ $+.66[\lambda \mathbf{I}]$	

Tableau A.6(c) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$0.41[\lambda\mathbf{DAD}'] + 0.48[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.11[\lambda\mathbf{I}]$	$[\lambda\mathbf{DAD}']$
crabes	$0.16[\lambda_k\mathbf{DAD}'] + 0.83[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$
haberman	$0.33[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.33[\lambda\mathbf{B}] + 0.34[\lambda\mathbf{I}]$	$[\lambda\mathbf{DA}_k\mathbf{D}']$
haemo	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
iris	$0.32[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.68[\lambda\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{DAD}']$
thyroïde	$0.30[\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.68[\lambda_k\mathbf{I}]$	$[\lambda_k\mathbf{B}]$
vin	$0.1[\lambda\mathbf{DA}_k\mathbf{D}'] + 0.25[\lambda_k\mathbf{DA}_k\mathbf{D}'] + 0.65[\lambda_k\mathbf{D}_k\mathbf{AD}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$

Tableau A.7 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *comité de méthodes* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$0.75[\lambda\mathbf{DAD}'] + 0.22[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
crabes	$0.48[\lambda\mathbf{DAD}'] + 0.52[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
haberman	$0.29[\lambda\mathbf{DAD}'] + 0.16[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.54[\lambda\mathbf{I}]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
haemo	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
iris	$0.53[\lambda\mathbf{DAD}'] + 0.37[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k] + 0.10[\lambda\mathbf{I}]$	$[\lambda\mathbf{DAD}']$
thyroïde	$0.15[\lambda\mathbf{DAD}'] + 0.85[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
vin	$0.31[\lambda\mathbf{DAD}'] + 0.69[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$

Tableau A.8 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *comité de méthodes* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$0.79[\lambda\mathbf{B}] + 0.21[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}]$
crabes	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{I}]$
haberman	$0.40[\lambda\mathbf{D}_k\mathbf{AD}'_k] + 0.60[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}_k]$
haemo	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
iris	$0.82[\lambda\mathbf{B}] + 0.18[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}]$
thyroïde	$0.73[\lambda\mathbf{B}] + 0.27[\lambda\mathbf{I}]$	$[\lambda\mathbf{B}]$

Tableau A.9 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *comité de méthodes* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
crabes	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$
haberman	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DA}_k\mathbf{D}']$
haemo	$.12[\lambda_k\mathbf{DAD}'] + .35[\lambda_k\mathbf{DA}_k\mathbf{D}']$ $+.14[\lambda_k\mathbf{D}_k\mathbf{AD}'_k] + .39[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
iris	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
thyroïde	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{B}]$
vin	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{D}_k\mathbf{AD}'_k]$

Tableau A.10 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
haberman	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
haemo	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
vin	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$

Tableau A.11 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{B}]$	$[\lambda \mathbf{B}]$
crabes	$[\lambda \mathbf{DAD}']$	$[\lambda \mathbf{I}]$
haberman	$[\lambda \mathbf{B}_k]$	$[\lambda \mathbf{B}_k]$
haemo	$.15[\lambda \mathbf{DAD}'] + .85[\lambda_k \mathbf{DAD}']$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$
thyroïde	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$
vin	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$

Tableau A.12 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
haberman	$.17[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}'] + .51[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + .31[\lambda_k \mathbf{B}_k]$	$[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$
haemo	$.14[\lambda \mathbf{DAD}'] + .3[\lambda_k \mathbf{DAD}'] + .13[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$ $+.3[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}'] + .13[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
iris	$.58[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k] + .42[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{B}]$
vin	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$

Tableau A.13 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances pénalisées* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
haberman	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
haemo	$.52[\lambda \mathbf{DAD}'] + .48[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
vin	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$

Tableau A.14 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances pénalisées* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k \mathbf{B}]$	$[\lambda \mathbf{B}]$
crabes	$[\lambda \mathbf{DAD}']$	$[\lambda \mathbf{I}]$
haberman	$[\lambda \mathbf{B}_k]$	$[\lambda \mathbf{B}_k]$
haemo	$.32[\lambda \mathbf{DAD}'] + .68[\lambda_k \mathbf{DAD}']$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$
thyroïde	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$
vin	$[\lambda \mathbf{B}]$	$[\lambda \mathbf{B}]$

Tableau A.15 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances pénalisées* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$.57[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}'] + .43[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
crabes	$.65[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k] + .35[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
haberman	$.42[\lambda \mathbf{B}_k] + .58[\lambda_k \mathbf{B}_k]$	$[\lambda \mathbf{D}\mathbf{A}_k \mathbf{D}']$
haemo	$.51[\lambda \mathbf{DAD}'] + .34[\lambda_k \mathbf{DAD}'] + .15[\lambda \mathbf{D}\mathbf{A}_k \mathbf{D}']$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{B}_k]$	$[\lambda_k \mathbf{B}]$
vin	$[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$

Tableau A.16 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances intégrées* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
crabes	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
haberman	$.36[\lambda\mathbf{DAD}'] + .64[\lambda_k\mathbf{DAD}']$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
haemo	$.83[\lambda\mathbf{DAD}'] + .17[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
iris	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda\mathbf{DAD}']$
thyroïde	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$
vin	$[\lambda\mathbf{DAD}']$	$[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}'_k]$

Tableau A.17 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances intégrées* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k\mathbf{B}]$	$[\lambda\mathbf{B}]$
crabes	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{I}]$
haberman	$[\lambda\mathbf{B}_k]$	$[\lambda\mathbf{B}_k]$
haemo	$.6[\lambda\mathbf{DAD}'] + .4[\lambda_k\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
iris	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
thyroïde	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
vin	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$

Tableau A.18 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *ratio des vraisemblances intégrées* avec les modèles les plus différents (II).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$.48[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}'] + .53[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
crabes	$.14[\lambda_k \mathbf{DAD}'] + .36[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k] + .48[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$
haberman	$.27[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}'] + .69[\lambda_k \mathbf{B}_k]$	$[\lambda \mathbf{D}\mathbf{A}_k \mathbf{D}']$
haemo	$.51[\lambda \mathbf{DAD}'] + .34[\lambda_k \mathbf{DAD}'] + .15[\lambda \mathbf{D}\mathbf{A}_k \mathbf{D}']$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{B}_k]$	$[\lambda_k \mathbf{B}]$
vin	$[\lambda_k \mathbf{D}\mathbf{A}_k \mathbf{D}']$	$[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$

Tableau A.19 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *empilement* avec les 14 modèles EDDA.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda \mathbf{DAD}']$	$[\lambda \mathbf{DAD}']$
crabes	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
haberman	$.36[\lambda \mathbf{DAD}'] + .64[\lambda_k \mathbf{DAD}']$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
haemo	$.83[\lambda \mathbf{DAD}'] + .17[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
iris	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda \mathbf{DAD}']$
thyroïde	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$
vin	$[\lambda \mathbf{DAD}']$	$[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$

Tableau A.20 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *empilement* avec les modèles les plus différents (I).

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
bupa	$[\lambda_k\mathbf{B}]$	$[\lambda\mathbf{B}]$
crabes	$[\lambda\mathbf{DAD}']$	$[\lambda\mathbf{I}]$
haberman	$[\lambda\mathbf{B}_k]$	$[\lambda\mathbf{B}_k]$
haemo	$.6[\lambda\mathbf{DAD}'] + .4[\lambda_k\mathbf{DAD}']$	$[\lambda\mathbf{DAD}']$
iris	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
thyroïde	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$
vin	$[\lambda\mathbf{B}]$	$[\lambda\mathbf{B}]$

Tableau A.21 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *empilement* avec les modèles les plus différents (II).

Jeu de données	groupe régressé	modèles contribuant à la combinaison et poids respectifs	modèle choisi
crabes	groupe 1	$[\lambda \mathbf{DAD'}]$	$[\lambda \mathbf{DAD'}]$
	groupe 2	$[\lambda \mathbf{DAD'}]$	
	groupe 3	$[\lambda \mathbf{DAD'}]$	
	groupe 4	$0.9[\lambda \mathbf{DAD'}] + 0.1(k - \text{ppv})$	
haberman	groupe 1	$0.51[\lambda \mathbf{DAD'}] + 0.18[\lambda \mathbf{I}] + 0.31(k - \text{ppv})$	$[\lambda \mathbf{DAD'}]$
	groupe 2	$0.43[\lambda \mathbf{DAD'}] + 0.27[\lambda \mathbf{I}] + 0.31(k - \text{ppv})$	
haemo	groupe 1	$[\lambda \mathbf{DAD'}]$	$[\lambda \mathbf{DAD'}]$
	groupe 2	$[\lambda \mathbf{DAD'}]$	
iris	groupe 1	$(k - \text{ppv})$	$[\lambda \mathbf{DAD'}]$, $(k - \text{ppv})$
	groupe 2	$[\lambda \mathbf{DAD'}]$	
	groupe 3	$[\lambda \mathbf{DAD'}]$	
thyroïde	groupe 1	$(k - \text{ppv})$	$(k - \text{ppv})$
	groupe 2	$(k - \text{ppv})$	
	groupe 3	$0.18[\lambda \mathbf{DAD'}] + 0.22[\lambda \mathbf{I}] + 0.60(k - \text{ppv})$	

Tableau A.22 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression separée* avec les méthodes $(k - \text{ppv})$, $[\lambda \mathbf{DAD'}]$ et $[\lambda \mathbf{I}]$.

Jeu de données	modèles contribuant à la combinaison et poids respectifs				modèle choisi
crabes					
groupe régressé	groupe 1	groupe 2	groupe 3	groupe 4	
groupe 1	$[\lambda\mathbf{DAD}']$				
groupe 2		$[\lambda\mathbf{DAD}']$			$[\lambda\mathbf{DAD}']$
groupe 3			$[\lambda\mathbf{DAD}']$		
groupe 4				$[\lambda\mathbf{DAD}']$	
haberman					
groupe régressé	groupe 1		groupe 2		
groupe 1	$.62[\lambda\mathbf{DAD}']$ $+.27(k-\text{ppv})$		$+.12[\lambda\mathbf{I}]$		$[\lambda\mathbf{DAD}']$
groupe 2			$.68[\lambda\mathbf{DAD}']$ $+.32(k-\text{ppv})$		
haemo					
groupe régressé	groupe 1		groupe 2		
groupe 1	$[\lambda\mathbf{DAD}']$				$[\lambda\mathbf{DAD}']$
groupe 2			$[\lambda\mathbf{DAD}']$		

Tableau A.23(a) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les méthodes $(k-\text{ppv})$, $[\lambda\mathbf{DAD}']$ et $[\lambda\mathbf{I}]$.

Jeu de données	modèles contribuant à la combinaison et poids respectifs			modèle choisi
iris				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$(k-\text{ppv})$			
groupe 2		$[\lambda\mathbf{DAD'}]$		$[\lambda\mathbf{DAD'}]$
groupe 3			$[\lambda\mathbf{DAD'}]$	
thyroïde				
groupe régressé	groupe 1	groupe 2	groupe 3	
groupe 1	$(k-\text{ppv})$			
groupe 2		$(k-\text{ppv})$		$(k-\text{ppv})$
groupe 3			$.20[\lambda\mathbf{DAD'}]$ $+.21[\lambda\mathbf{I}]$ $+.59(k-\text{ppv})$	

Tableau A.23(b) : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *régression vectorisée* avec les méthodes $(k-\text{ppv})$, $[\lambda\mathbf{DAD'}]$ et $[\lambda\mathbf{I}]$.

Jeu de données	modèles contribuant à la combinaison et poids respectifs	modèle EDDA
crabes	$0.90[\lambda\mathbf{DAD'}] + 0.10(k - \text{ppv})$	$[\lambda\mathbf{DAD'}]$
haberman	$0.43[\lambda\mathbf{DAD'}] + 0.26[\lambda\mathbf{I}] + 0.31(k - \text{ppv})$	$[\lambda\mathbf{DAD'}]$
haemo	$[\lambda\mathbf{DAD'}]$	$[\lambda\mathbf{DAD'}]$
iris	$0.90[\lambda\mathbf{DAD'}] + 0.10(k - \text{ppv})$	$[\lambda\mathbf{DAD'}]$
thyroïde	$0.1[\lambda\mathbf{DAD'}] + 0.90(k - \text{ppv})$	$(k - \text{ppv})$

Tableau A.24 : Modèles contribuant à la combinaison et poids respectifs pour chaque jeu de données par la méthode de combinaison *comité de méthodes* avec les méthodes $(k - \text{ppv})$, $[\lambda\mathbf{DAD'}]$ et $[\lambda\mathbf{I}]$.

Références

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control AU-19*, 716–22.

Arinze, B., S. Kim, and M. Anandarajan (1997). Combining and selecting forecasting models using rule based induction. *Computers Ops. Res. 24*(5), 423–33.

Bensmail, H. and G. Celeux (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association 91*, 1743–48.

Berger, J. (1980). *Statistical Decision Theory.* New York : Sringer-Verlag.

Berger, J. and M. Bock (1976). Combining independent normal mean estimation problems with unknown variances. *Annals of Statistics 4*(3), 642–48.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Londres : Oxford University Press.

Blake, C. and C. Merz (1998). UCI repository of machine learning databases. http ://www.ics.uci.edu/~mlearn/MLRepository.html.University of California, Irvine, Dept. of Information and Computer Sciences.

Breiman, L. (1994). Heuristics of instability in model selection. Technical report, Department of Statistics, University of California.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*, 123–40.

Breiman, L. (1996b). Stacked regressions. *Machine Learning 24*, 49–64.

Breiman, L. and J. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistic Society 59*(1), 3–54.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Monterey : Wadsworth and Brooks Cole.

Brito, I. and G. Celeux (1999). Combining edda's models. In H. Bacela-Nicolau, F. Nicolau, and J. Jansen (Eds.), *Applied Stochastic Models and Data Analysis*, Lisboa. Instituto Nacional de Estatística.

Brito, I. and G. Celeux (2000). Discriminant analysis by hierarchical coupling in edda context. In J. Jansen (Ed.), *Proceedings of IFCS- 2000*, Namur. Springer-Verlag.

Brito, I. and G. Celeux (2001). Combinação de modelos gaussianos em análise discriminante : algumas notas. In P. Sousa (Ed.), *Proceedings of Joclad 2001*, Porto. Instituto Nacional de Estatística.

Celeux, G. (1991). *Analyse discriminante sur variables continues*. Paris : Collection Didactique INRIA.

Celeux, G., I. Brito, and A. Ferreira (1999). Combining models in discriminant analysis through committees of methods. Inria Rhône-Alpes, IS2.

Celeux, G. and J.-P. Nakache (1994). *Analyse discriminante sur variables qualilatives*. Paris : Polytechnica.

Cho, S. and J. Kim (1995a). Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Transactions on Systems, Man and Cybernetics 25*(2), 380–4.

Cho, S. and J. Kim (1995b). Multiple network fusion using fuzzy logic. *IEEE Transactions on Neural Networks 6*(2), 497–501.

Clemen, R. (1989). Combining forecasts : a review and annotated bibliography. *International Journal of Forecasting 5*, 559–83.

Deming, W. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematics and Statistics*, 427–44.

Dietterich, T. and G. Bakiri (1995). Error-correcting output codes : a general method for improving multiclass inductive learning programs. In *Proceedings of the 9th National Conference on Artificial Intelligence*, Anaheim, California. AAAI Press.

Efron, B. and C. Morris (1973). Combining possibly related estimation problems(with discussion). *Journal of the Royal Statistical Society, Ser. B 35*, 379–421.

Ferreira, A. (2001). *Combinação de modelos em análise discriminante sobre variáveis qualitativas.* Ph. D. thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 179–88.

Freund, Y. and R. Schapire (1996). Experiments with a new boosting algoritm. In J. Cowan, G. Tesauro, and J. Alspector (Eds.), *Proceedings of the 13th International Conference on Machine Learning*, San Mateo, California, pp. 148–56. Morgan Kaufman.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association 84*, 165.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics 19*, 1–141.

Friedman, J. (1996). Another approach to polychotomous classification. Technical report, Stanford University.

Granger, C. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of Forecasting 3*, 197–204.

Green, E. and W. Strawderman (1991). A james-stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association 86*(416), 1001–6.

Habbema, J., J. Hermans, and K. van den Broek (1974). A stepwise discriminante analysis program using density estimation. In *COMPSTAT 1974, Proceedings in Computational Statistics*, pp. 101–10. Physica-Verlag.

Hansen, L. and P. Salamon (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(10), 993–1001.

Hashem, S. and B. Schmeiser (1995). Improving model accuracy using optimal linear combination of trained neural networks. *IEEE Transactions on Neural Networks 6*(3), 792–4.

Hastie, T. and R. Tibshirani (1993). Varying coefficient models(with discussion). *Journal of the Royal Statistical Society, Serie B 55*, 757–96.

Hastie, T. and R. Tibshirani (1996). Classification by pairwise coupling. Technical report, University of Toronto.

Hoerl, A.and Kennard, R. (1970). Ridge regression : biased estimation for nonorthogonal problems. *Techometrics 8*, 27–51.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging : a tutorial. *Statistical Science 14*(4), 179–91.

Jacobs, R., M. Jordan, S. Nowlan, and G. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*(1), 79–87.

Jordan, M. and R. Jacobs (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation 6*, 181–214.

Judge, G., W. Griffiths, C. Hill, and T. Lee (1980). *The Theory and Practice of Econometrics*. New York : Wiley.

Kass, R. and L. Wasserman (1995). A reference bayesian test for nested hypotheses with large samples. *Journal of the American Statistical Society 90*, 928–34.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. New York : Wiley.

Kivinen, J. and M. Warmuth (1997). Exponentiated gradient descent versus gradient descend for linear predictors. *Information and Computation 132*(1), 1–63.

Krogh, A. and J. Vedelsby (1995). Neural network ensembles, cross validation and active learning. In G. Tesauro, D. Touretzky, and T. Leen (Eds.), *Advances in neural information Processing Systems 7*, Cambridge, Massachusetts, pp. 231–8. MIT Press.

Lawson and Hanson (1974). *Solving Least Squares Problems*. Londres : Prentice Hall.

Lebart, L., A. Morineau, and M. Piron (1995). *Statistique Exploratoire Multidimensionnelle*. Paris : Dunod.

LeBlanc, M. and R. Tibshirani (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association 91*(436), 1641–50.

Levin, A., T. Leen, and J. Moody (1994). Fast pruning using principal components. In G. Cowan, J.and Tesauro and J. Alspector (Eds.), *Advances in neural information processing systems 6*, San Mateo, California, pp. 35–42. Morgan Kaufman.

Madigan, D., A. Raftery, and J. Hoeting (1994). Model selection and accounting for model uncertainty in graphical models using occam's window.

Journal of the American Statistical Association 89, 1535–46.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winckler (1982). The accuracy of extrapolation(time series) methods : results of a forecasting competition. *Journal of Forecasting 1*, 111–53.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York : Wiley.

Meir, R. (1995). Bias, variance and the combination of least squares estimators. In D. Tesauro, G.and Touretzky and T. Leen (Eds.), *Advances in neural information processing systems 7*, Cambridge, Massachusetts, pp. 295–302. MIT Press.

Merz, C. (1999). Using correspondence analysis to combine classifiers. *Machine Learning 36*, 59–83.

Merz, C. and M. Pazzani (1999). A principal components approach to combining regressions estimates. *Machine Learning 36*, 9–32.

Opitz, D. and J. Shavlik (1996). Generating accurate and diverse members of a neural-network ensemble. In D. Touretzky, M. Mozer, and M. Hasselmo (Eds.), *Advances in neural information processing systems 8*, Cambridge, Massachusetts, pp. 535–41. MIT Press.

Perrone, M. and L. Cooper (1973). When networks disagree : ensemble methods for hybrid neural networks. In R. Mammone (Ed.), *Artificial neural networks for speech and vision*, Londres, pp. 126–42. Chapman & Hall.

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika 83*(2), 251–266.

Raftery, A., D. Madigan, and J. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Rao, J. and K. Subrathmaniam (1971). Combining independent estimators and estimation in linear regression with unequal variances. *Biometrics 27*, 971–90.

Ripley, B. (1996). *Pattern Recognition and Neural Networks.* Cambridge : Cambridge University Press.

Rogova, G. (1994). combining the results of several neural network classifiers. *Neural Networks 7*(5), 777–81.

Rubin, D. and S. Weisberg (1975). The variance of a linear combination of independent estimators using estimated weights. *Biometrika 62*, 708–9.

Sanders, R. and L. Ritzman (1990). Improving short-term forecasts. *Omega 18*(4), 365–73.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning 5*, 197–227.

Schnaars, S. (1984). Situational factors affecting forecast accuracy. *Journal of Marketing Res. 21*, 290–7.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–4.

Smyth, P. and D. Wolpert (1999). Linear combining density estimators. *Machine Learning 36*, 59–83.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Ser. B 36*, 111–47.

Szabo, F.and Rubin, C., F. Esteve, S. Grand, M. Décops, V. Lefournier, J.-F. Le Bas, and C. Rémy (2000). A new approach for analyzing proton tumors : nosologic images. *Nature Medecine 6*(11), 1287–89.

Szabo, F. (2001). *Caractérisation par IRM de paramètres métaboliques tissulaires en pathologie tumorale cérébrale chez l'homme.* Ph. D. thesis, Université Joseph Fourier - Grenoble I.

Taplin, R. (1993). Robust likelihood calculation for time series. *Journal of the Royal Statistical Society B 55*, 829–36.

Tresp, V. and M. Taniguchi (1995). Combining estimators using non-costant weighting functions. In G. Tesauro, D. Touretzky, and T. Leen (Eds.), *Advances in neural information processing systems 7*, Cambridge, Massachusetts, pp. 295–302. MIT Press.

van der Merwe, A. and J. Zidek (1980). Multivariate regression analysis and canonical variates. *Can. Journal of Statistics 8*, 27–39.

Venables, W. and B. Ripley (1994). http ://www.stats.ox.ac.uk/~pub/Sbook. University of Oxford, Dept. of Statistics.

Winckler, R. (1989). Combining forecasts : a philosophical basis and some current issues. *International Journal of Forecasting 5*, 605–9.

Wolpert, D. (1992). Stacked generalization. *Neural Networks 5*, 241–59.