Analyse Discriminante de Haute Dimension

Charles Bouveyron, Stéphane Girard, Cordelia Schmid

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Analyse Discriminante de Haute Dimension

Charles Bouveyron — Stéphane Girard — Cordelia Schmid

N° 5470

Janvier 2005

Thèmes COG et NUM

ISSN 0249-6399 ISRN INRIA/RR--5470--FR

Analyse Discriminante de Haute Dimension

Charles Bouveyron [*] [†] , Stéphane Girard [*] , Cordelia Schmid [†]

Thèmes COG et NUM —Systèmes cognitifs et Systèmes numériques
Projets Mistis et Lear

Rapport de recherche n° 5470 —Janvier 2005 — 43 pages

Résumé : Nous proposons une nouvelle méthode d'Analyse Discriminante, nommée Analyse Discriminante de Haute Dimension (HDDA), adaptée aux données de grande dimension. Notre approche est basée sur l'hypothèse que les données de grande dimension vivent dans des sous-espaces dont la dimension intrinsèque est inférieure à la dimension de l'espace. Pour ce faire, nous réduisons tour à tour la dimension des données de chaque classe et nous régularisons la matrice de covariance de la classe afin d'adapter le modèle gaussien de l'analyse discriminante à ce type de données. Cette stratégie conduit à une nouvelle règle de décision qui comporte un certain nombre de cas particuliers ayant une interprétation géométrique. Nous présentons les résultats de la mise en œuvre de cette méthode de classification multi-classes sur des données artificielles et réelles. En particulier, nous appliquons notre méthode à la reconnaissance de classe d'objets dans des images naturelles.

Mots-clés : Analyse Discriminante, réduction de dimension, régularisation

This work was supported by the French department of Research through the *ACI Masse de données* (MoViStaR project)

[*] INRIA Rhône-Alpes, projet Mistis & LMC-IMAG, laboratoire SMS
[†] INRIA Rhône-Alpes, projet Lear

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier (France)
Téléphone : +33 4 76 61 52 00 —Télécopie +33 4 76 61 52 52

High Dimensional Discriminant Analysis

Abstract: We propose a new method for discriminant analysis, called High Dimensional Discriminant Analysis (HHDA). Our approach is based on the assumption that high dimensional data live in different subspaces with low dimensionality. Thus, HDDA reduces the dimension for each class independently and regularizes class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. This regularization is achieved by assuming that classes are spherical in their eigenspace. HDDA is applied to recognize objects in natural images and its performances are compared to classical classification methods.

Key-words: Discriminant analysis, dimension reduction, regularization

Table des matières

Notations

C_i : $i^{\text{ème}}$ classe connue *a priori*,
k : nombre total de classes,
p : dimension de l'espace total,
x : vecteur de $\mathbb{R}^p$,
δ^* : règle de décision de Bayes,
δ^+ : règle de décision de l'HDDA,
π_i : proportion de la classe C_i,
μ_i : moyenne de la classe C_i,
Σ_i : matrice de variance de la classe C_i,
Σ : matrice de variance commune à toutes les classes,
Σ_{totale} : matrice de variance de l'ensemble des données,
$\mathcal{B}_i$: base des vecteurs propres de Σ_i,
Q_i : matrice de passage de la base canonique de $\mathbb{R}^p$ à $\mathcal{B}_i$,
Δ_i : matrice de variance de la classe C_i dans $\mathcal{B}_i$,
$\mathbb{E}_i$: espace propre dans lequel vivent les données de la classe C_i,
$P_i(.)$: opérateur de projection sur $\mathbb{E}_i$,
d_i : dimension de l'espace $\mathbb{E}_i$,
$\mathbb{E}_i^{\perp}$: espace supplémentaire de $\mathbb{E}_i$,
$P_i^{\perp}(.)$: opérateur de projection sur $\mathbb{E}_i^{\perp}$.

Glossaire

HDDA : Analyse Discriminante de Haute Dimension,
RDA : Analyse Discriminante Régularisée,
EDDA : Analyse Discriminante à Décomposition Spectrale,
LDA : Analyse Discriminante Linéaire,
QDA : Analyse Discriminante Quadratique,
FDA : Analyse Discriminante Factorielle,
SVM : Machines à Vecteurs Supports.

1 Introduction

L'apprentissage statistique est actuellement confronté à des données qui sont représentées dans des espaces de très grande dimension. On retrouve, par exemple, cette caractéristique pour les données issues de la biologie (puces ADN), de l'analyse textuelle ou de la vision par ordinateur. Cependant, la classification des données de grande dimension est un problème très difficile. En effet, les recherches menées ces dernières années ont montré que le traitement de ce type de données révèle des phénomènes très différents de ce que l'on connaît dans les espaces usuels. En particulier, dans des espaces de grande dimension, les performances des méthodes d'apprentissage statistique souffrent du phénomène bien connu du *fléau de la dimension* [3]. Le *phénomène de l'espace vide* [25], dû au fait que le volume total de l'espace croît exponentiellement en fonction de la dimension [27, 8], nous permet de supposer que les données vivent dans des espaces de dimension intrinsèque plus faible. L'approche que nous exposons dans ce rapport se propose de reformuler le modèle statistique de l'Analyse Discriminante afin de prendre en compte les spécificités des données de grande dimension.
Nous rappellerons tout d'abord, au paragraphe 2, le cadre général de l'Analyse Discriminante et le problème de la classification. Nous présenterons ensuite, au paragraphe 3, le modèle statistique adapté aux données de grande dimension et la construction de la règle de décision de notre méthode appelée Analyse Discriminante de Haute Dimension[1]. Nous détaillerons également, au paragraphe 4, les cas particuliers issus de notre nouvelle règle de décision. Enfin, aux paragraphes 6 et 7, notre méthode sera mise en œuvre et comparée à des méthodes de références sur des données synthétiques et réelles.

2 Cadre général de l'Analyse Discriminante

Le problème de l'Analyse Discriminante, également connue sous le nom de *classification supervisée*, est de prédire l'appartenance d'un individu x à une classe parmi k. On distingue classiquement deux objectifs principaux en Analyse Discriminante :

(i) *descriptif* : l'aspect *descriptif* vise à trouver une représentation qui permette l'interprétation des groupes grâce aux variables explicatives.

(ii) *décisionnel* : dans ce cas, on cherche à définir la *bonne* affectation d'un nouvel individu dont on ne connaît que les valeurs des variables explicatives. Cet aspect est particulièrement apprécié dans des domaines où l'aspect diagnostic est essentiel.

Dans ce rapport, nous nous intéresserons plus particulièrement à l'aspect *décisionnel* qui est le plus important et souvent le plus délicat. Nous allons tout d'abord rappeler le modèle probabiliste de la discrimination avant de décrire les principales méthodes paramétriques d'Analyse Discriminante.

2.1 Le problème de la discrimination

Afin de prédire l'appartenance d'un individu x, décrit par p variables explicatives, à une classe parmi k classes C_1, ..., C_k définies *a priori*, nous disposons d'un ensemble d'*apprentissage* :

$$\mathcal{A} = \{(x_1, c_1), ..., (x_n, c_n),\ x_i \in \mathbb{R}^p,\ c_i \in \{1, ..., k\}\},$$

où le vecteur x_i contient les valeurs prises par le $i^{\text{ème}}$ individu sur les p variables explicatives et c_i indique le numéro de la classe à laquelle x_i appartient. Nous allons utiliser l'échantillon A pour construire une règle de décision δ qui associe à tout vecteur de $\mathbb{R}^p$ une des k classes :

$$\begin{array}{rccl} \delta\ :\ \mathbb{R}^p & \longrightarrow & \{1, ..., k\}, \\ x & \longmapsto & c. \end{array}$$

[1] *High Dimensional Discriminant Analysis* (HDDA) en anglais.

2.2 La règle de décision de Bayes

Les méthodes paramétriques d'Analyse Discriminante font une hypothèse de normalité des classes. Les densités de probabilité des variables explicatives conditionnellement aux classes $p(x|C_i)$, $\forall i = 1, ..., k$ sont supposées celles de lois normales $\mathcal{N}(\mu_i, \Sigma_i)$ de moyennes μ_i et de matrice de variance Σ_i :

$$p(x|C_i) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_i)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)\right). \tag{2.1}$$

La règle de décision optimale δ^*, dite *règle de Bayes*, affecte le point x à la classe qui a la probabilité *a posteriori* maximum. La règle de décision δ^* prend la forme suivante :

$$x \in C_{i^*}, \text{ si } i^* = \underset{i=1,...,k}{\operatorname{argmax}}\{p(C_i|x)\}. \tag{2.2}$$

La formule de Bayes permet d'obtenir la probabilité que l'individu x provienne de la classe C_i :

$$p(C_i|x) = \frac{p(x|C_i)\pi_i}{p(x)}, \tag{2.3}$$

où π_i est la probabilité *a priori* de la classe C_i et où $p(x) = \sum_{i=1}^{k} p(x|C_i)$. Par conséquent et comme les dénominateurs de l'équation (2.3) sont communs pour chacune des k classes, la règle de Bayes consiste donc à affecter x à la classe C_{i^*} si :

$$i^* = \underset{i=1,...,k}{\operatorname{argmax}}\{\pi_i\, p(x|C_i)\}. \tag{2.4}$$

Définition 2.1. Afin de faciliter l'écriture des règles de décision par la suite, nous allons définir la fonction de coût K_i conditionnellement à la classe C_i, $\forall i = 1, ..., k$, de la façon suivante :

$$\begin{aligned} K_i \,:\, \mathbb{R}^p &\longrightarrow \mathbb{R}, \\ x &\longmapsto -2\log(\pi_i p(x|C_i)). \end{aligned}$$

Avec cette notation, la règle de Bayes consiste donc à affecter x à la classe C_{i^*} si :

$$i^* = \underset{i=1,...,k}{\operatorname{argmin}}\{K_i(x)\}.$$

2.3 Les principales méthodes d'Analyse Discriminante

Dans cette section, nous allons décrire brièvement les principales méthodes paramétriques d'Analyse Discriminante. Pour plus de détails, on pourra consulter [7] et [23, chap. 18].

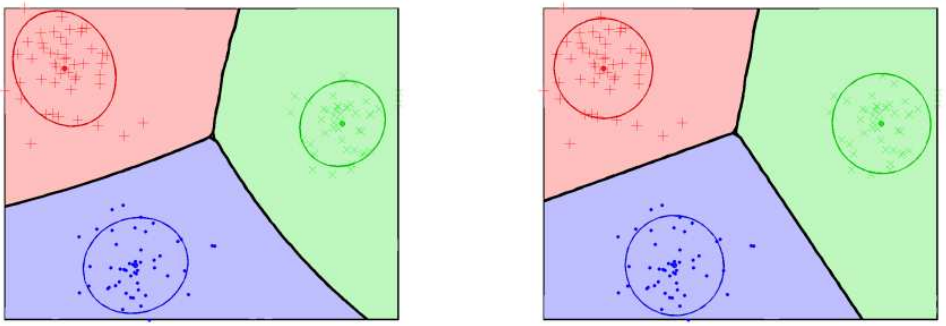
Analyse Discriminante Quadratique (QDA) Avec les hypothèses précédentes, la règle de décision δ^* revient à minimiser la fonction de coût :

$$K_i(x) = (x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) + \log(\det \Sigma_i) - 2\log(\pi_i) + C^{te}, \tag{2.5}$$

où la constante représente une quantité commune à toutes les classes et n'intervient donc pas dans la règle de décision. Lorsque les Σ_i sont différentes, cette règle réalise des séparations quadratiques entre les classes (voir Fig. 2.1-a). En pratique, cette méthode est pénalisée par l'estimation des nombreux paramètres quand la dimension p devient grande.

Analyse Discriminante Quadratique à classes sphériques (QDAs) Afin de pallier la limitation précédente, on peut faire l'hypothèse que les matrices de variance des classes sont proportionnelles à l'identité $\Sigma_i = \sigma_i^2 Id$, c'est à dire que les classes sont de forme sphérique. Avec ces hypothèses, la règle de décision δ^* revient alors à minimiser la fonction de coût :

$$K_i(x) = \frac{1}{\sigma_i^2}\|x-\mu_i\|^2 + 2p\log(\sigma_i) - 2\log(\pi_i) + C^{te}. \tag{2.6}$$



(a) (b)

FIG. 2.1 – Frontières de décision de (a) l'Analyse Discriminante Quadratique et de (b) l'Analyse Discriminante Linéaire sur un même jeu de données.

Analyse Discriminante Linéaire (LDA) Si, par rapport à l'Analyse Discriminante Quadratique, on fait l'hypothèse supplémentaire d'égalité des matrices de variances, *i.e.* $\forall i = 1, ..., k,\ \Sigma_i = \Sigma$, on se place alors dans le cadre de l'Analyse Discriminante Linéaire qui doit son nom au fait qu'elle réalise des séparations linéaires entre les classes (voir Fig. 2.1-b). En effet, on peut alors éliminer les termes $\log(\det \Sigma_i)$, qui sont alors constants, et $x^t\Sigma^{-1}x$ qui ne dépend pas de la classe. Alors, la règle de décision δ^* revient à minimiser la fonction de coût :

$$K_i(x) = (x-\mu_i)^t\Sigma^{-1}(x-\mu_i) - 2\log(\pi_i) + C^{te}, \quad (2.7)$$

$$= \mu_i^t\Sigma^{-1}\mu_i - 2\mu_i^t\Sigma^{-1}x - 2\log(\pi_i) + C^{te}. \quad (2.8)$$

En pratique, l'Analyse Discriminante Linéaire est fréquemment utilisée car elle offre un bon compromis entre pertinence et complexité. En effet, elle fournit des résultats robustes aux fluctuations sur les hypothèses de normalité des classes et d'égalité des matrices de variance. Pour ces raisons, elle doit être considérée comme une méthode de référence.

Analyse Discriminante Linéaire à classes sphériques (LDAs) Si de plus, on suppose que $\Sigma = \sigma^2 Id$, c'est à dire que les classes sont de même forme sphérique, alors la règle de décision δ^* revient à minimiser la fonction de coût :

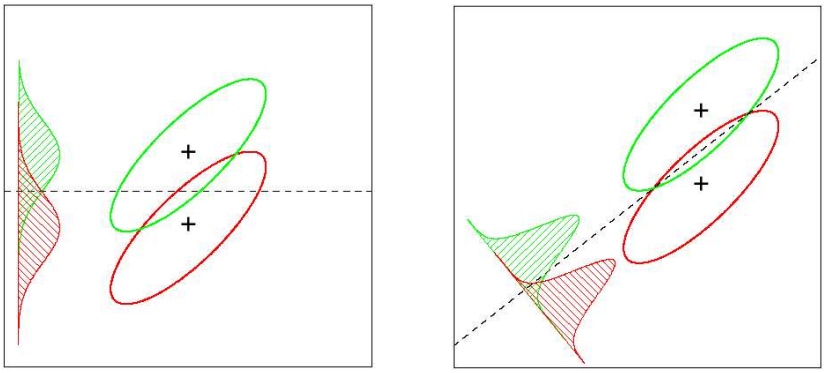
$$K_i(x) = \frac{1}{\sigma^2}\|x-\mu_i\|^2 - 2\log(\pi_i) + C^{te}. \quad (2.9)$$

Règle géométrique de l'Analyse Discriminante Linéaire (LDAgéo) Si l'on fait l'hypothèse supplémentaire que $\Sigma = Id$ et que les proportions sont égales $\pi_i = \pi_*$, alors on obtient la règle géométrique de l'Analyse Discriminante Linéaire qui consiste à minimiser la fonction de coût :

$$K_i(x) = \|x-\mu_i\|^2 + C^{te}, \quad (2.10)$$

qui affecte le point x à la classe dont il est le plus proche de la moyenne. Ce classifieur a été baptisé *nearest-means classifer* par Friedman [13]. Toutefois, cette règle simple conduit à des erreurs d'affectation quand la dispersion des classes est trop différente.

Analyse Factorielle Discriminante (FDA) L'Analyse Factorielle Discriminante combine une étape qui relève de la réduction de dimension et une étape de discrimination. En effet, effectuer une Analyse Factorielle Discriminante consiste à projeter les données de $\mathbb{R}^p$ sur les $d = (k-1)$ axes discriminants qui maximisent le rapport de la variance inter-classe et de la variance intra-classe, puis d'apprendre la règle de



(a) (b)

FIG. 2.2 – L'axe principal de la figure (a) ne permet pas de discriminer efficacement les deux groupes alors que celui de la figure (b) possède un bon pouvoir discriminant.

décision δ^* sur les données projetées. Nous avons donc besoin de définir les matrices de variance inter et intra-classe. La matrice de variance inter-classe est définie par :

$$B = \sum_{i=1}^{k} \pi_i (\mu_i - \mu)(\mu_i - \mu)^t,$$

où $\mu = \sum_{i=1}^{k} \pi_i \mu_i$. D'autre part, la matrice de variance intra-classe est définie par :

$$W = \sum_{i=1}^{k} \pi_i \Sigma_i.$$

Notons également que le théorème de Huyghens nous permet d'obtenir la relation suivante qui lie les matrices de variance inter et intra-classe à la matrice de variance totale :

$$\Sigma_{totale} = B + W.$$

Nous souhaitons trouver une représentation des données qui permettent de discriminer les groupes le mieux possible. Pour ce faire, il faut que les projections des k centres de gravité soient le plus séparées possible, tandis que les données de chaque classe doivent se projeter de façon groupée autour du centre de gravité de leur classe. Nous recherchons donc une représentation des données qui maximise la variance inter-classe et qui minimise la variance intra-classe. Avec les notations et résultats précédents, les axes de la projection recherchée satisfont le problème d'optimisation suivant :

$$\max_u \frac{u'Bu}{u'\Sigma_{totale}u}. \tag{2.11}$$

On sait que ce maximum est atteint pour u vecteur propre de $\Sigma_{totale}^{-1}B$ associé à sa plus grande valeur propre. La figure 2.2 illustre le choix d'un axe de projection permettant de discriminer au mieux les classes. Une fois la projection déterminée, on peut alors effectuer une Analyse Discriminante Linéaire (ou Quadratique). Cette stratégie qui combine réduction de dimension et discrimination est souvent profitable car les données de chaque classe n'occupent en général pas la totalité de l'espace et cela permet de réduire le nombre de paramètres à estimer. Cette méthode se révèle relativement efficace sur des données de grande dimension.

2.4 Régularisation de l'Analyse Discriminante

Comme nous l'avons dit, l'Analyse Discriminante Linéaire peut être considérée comme une méthode de référence du fait de sa robustesse. Toutefois, cette propriété de robustesse n'est plus vérifiée quand la

taille de l'échantillon devient trop petit devant la dimension de l'espace. Cette remarque est encore plus vraie en ce qui concerne l'Analyse Discriminante Quadratique. Au début des années 1990, des méthodes dites d'Analyse Discriminante régularisée ont vues le jour, ayant comme but de stabiliser les résultats de l'Analyse Discriminante dans ce cas limite. On pourra consulter [21] pour une synthèse sur le sujet.

L'Analyse Discriminante Régularisée (RDA) Friedman [13] propose de faire dépendre l'estimation des matrices de covariance des groupes de deux paramètres de régularisation λ et γ de la façon suivante :

$$\hat{\Sigma_i}(\lambda,\gamma) = (1-\gamma)\hat{\Sigma_i}(\lambda) + \gamma\left(\frac{tr(\hat{\Sigma_i}(\lambda))}{p}\right)Id,$$

où :

$$\hat{\Sigma_i}(\lambda) = \frac{(1-\lambda)(n_i-1)\hat{\Sigma_i} + \lambda(n-k)\hat{\Sigma}}{(1-\lambda)(n_i-1)+\lambda(n-k)}.$$

Le paramètre de complexité $\lambda \in [0,1]$ contrôle la contribution des estimateurs $\hat{\Sigma_i}$ et $\hat{\Sigma}$, qui sont respectivement les estimateurs de Σ_i et Σ définis au paragraphe 2.3 (LDA) . D'autre part, le paramètre $\gamma \in [0,1]$ contrôle l'estimation des valeurs propres des matrices de covariance. Ainsi, l'Analyse Discriminante Régularisée engendre une règle de décision qui « varie » entre l'Analyse Discriminante Linéaire et l'Analyse Discriminante Quadratique. Une application de cette méthode à la reconnaissance de visage est proposée dans [22].

L'Analyse Discriminante à Décomposition Spectrale (EDDA) L'EDDA (*Eigenvalue Decomposition Discriminant Analysis*) [4] propose une reparamétrisation des matrices de covariance des groupes afin d'éviter le recours aux paramètres de régularisation de la RDA. Cette méthode repose sur la décomposition spectrale suivante des matrices de covariance :

$$\Sigma_i = \lambda_i D_i A_i D_i^t,$$

où D_i est la matrice des vecteurs propres de Σ_i, A_i est une matrice diagonale contenant les valeurs propres normalisées et ordonnées de Σ_i et $\lambda_i = |\Sigma_i|^{1/p}$. Les quantités λ_i, D_i et A_i contrôlent respectivement le volume, l'orientation et la forme de la distribution de la classe C_i. En faisant varier ou non ces trois quantités, les auteurs mettent en évidence 14 modèles particuliers. L'EDDA choisit, par validation croisée, parmi ces modèles celui qui possède le plus petit taux d'erreur. Cette reparamétrisation permet, dans un certain nombre de modèles particuliers, que l'estimation des matrices de covariance ne se fasse qu'avec un nombre limité de paramètres à estimer. Cependant, ces estimateurs n'ont pas toujours une forme explicite et sont estimés par des méthodes itératives.

3 Analyse Discriminante de Haute Dimension

Les méthodes classiques d'Analyse Discriminante, présentées au paragraphe précédent, fournissent généralement des résultats satisfaisants pour des données de petite dimension et possèdent l'avantage d'avoir un fondement statistique. Cependant, ces méthodes sont pénalisées en haute dimension car la taille de l'échantillon d'apprentissage devient trop petit devant la dimension de l'espace et les paramètres ne sont plus estimés correctement. En particulier, les matrices de covariance des classes Σ_i ne sont alors pas bien estimées et peuvent devenir singulières.
Le phénomène de *l'espace vide* nous permet de supposer que les données de haute dimension vivent dans des sous-espaces différents et de dimension intrinsèque inférieure à la dimension de l'espace. Afin d'adapter le modèle gaussien de l'Analyse Discriminante aux données de grande dimension et de limiter le nombre de paramètres à estimer, nous proposons de projeter les données de chaque classe dans leur espace propre que nous décomposerons en deux sous-espaces supplémentaires de dimension inférieure et de faire l'hypothèse que les classes sont sphériques dans ces sous-espaces. Cette hypothèse de sphéricité se traduit par le fait que les nouvelles matrices de covariance des classes n'ont que deux valeurs propres différentes. De manière similaire à l'EDDA, notre méthode comportera plusieurs cas particuliers possédant, pour certains, des interprétations géométriques.

3.1 Définitions et hypothèses

De manière similaire à l'Analyse Discriminante classique, nous supposons que les densités de probabilité f_i des variables explicatives conditionnellement aux classes C_i, $\forall i = 1, ..., k$ sont normales $\mathcal{N}(\mu_i, \Delta_i)$ de moyennes μ_i et de matrice de variance Σ_i. Nous allons également définir les différents espaces dans lesquels nous travaillerons.

Définition 3.1. On appelle Q_i, $\forall i = 1, ..., k$, la matrice orthogonale des vecteurs propres de la matrice de variance Σ_i. On définit $\mathcal{B}_i$ comme étant la base de $\mathbb{R}^p$ composée des vecteurs propres de Σ_i. Ainsi, on peut définir dans $\mathcal{B}_i$ la matrice de covariance Δ_i de la classe C_i de la façon suivante :

$$\forall i = 1, ..., k, \; \Delta_i = Q_i^t \, \Sigma_i \, Q_i,$$

avec $Q_i^t Q_i = Id$.

Ainsi, dans la base $\mathcal{B}_i$ des vecteurs propres de Σ_i, la matrice Δ_i est diagonale et constituée des valeurs propres de Σ_i. Nous supposons de plus que la matrice Δ_i n'a que deux valeurs propres distinctes $a_i > b_i$:

$$\Delta_i = \begin{pmatrix} \boxed{\begin{matrix} a_i & & 0 \\ & \ddots & \\ 0 & & a_i \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{matrix}} \end{pmatrix} \begin{matrix} \} \, d_i \\ \} \, (p - d_i) \end{matrix} \tag{3.1}$$

Remarque 1. Nous supposerons dans ce document que $\forall i = 1, ..., k$, $d_i < p$ car nous nous trouvons sinon dans le cas de l'Analyse Discriminante classique avec l'hypothèse supplémentaire que les classes sont sphériques.

Définition 3.2. On définit à présent $\mathbb{E}_i$ comme étant l'espace affine engendré par les vecteurs propres associés à la valeur propre a_i et passant par le barycentre μ_i de la classe C_i (voir Fig. 3.1). On définit également $\mathbb{E}_i^\perp$ l'espace supplémentaire de $\mathbb{E}_i$ dans $\mathbb{R}^p$, *i.e.* $\mathbb{E}_i^\perp$ est l'espace engendré par les vecteurs propres associés à la valeur propre b_i et passant μ_i.

Par conséquent, la classe est de forme sphérique dans les espaces $\mathbb{E}_i$ et $\mathbb{E}_i^\perp$. Nous reviendrons par la suite sur la méthode d'estimation de d_i et des espaces $\mathbb{E}_i$ et $\mathbb{E}_i^\perp$ (voir paragraphe 5.5). Afin de construire la nouvelle règle de décision, nous avons besoin de définir, $\forall i = 1, ..., k$, les opérateurs de projection sur $\mathbb{E}_i$ et sur $\mathbb{E}_i^\perp$.

Définition 3.3. On définit, $\forall i = 1, ..., k$, l'opérateur P_i de projection d'un élément $x \in \mathbb{R}^p$ sur $\mathbb{E}_i$:

$$P_i \; : \; x \mapsto \tilde{Q}_i \tilde{Q}_i^{\,t} (x - \mu_i) + \mu_i,$$

où la matrice $\tilde{Q}_i$, de dimension $p \times p$, est composée des d_i premières colonnes de Q_i complétée par des 0. De même, on définit l'opérateur $P_i^\perp$ de projection d'un élément $x \in \mathbb{R}^p$ sur $\mathbb{E}_i^\perp$:

$$P_i^\perp \; : \; x \mapsto (Q_i - \tilde{Q}_i)(Q_i - \tilde{Q}_i)^t (x - \mu_i) + \mu_i,$$

où la matrice $(Q_i - \tilde{Q}_i)$, de dimension $p \times p$, est composée des $(p - d_i)$ dernières colonnes de Q_i complétée par des 0.

Remarque 2. On rappelle que, $\forall i = 1, ..., k$, le barycentre μ_i est invariant par l'opérateur P_i car $\mu_i \in \mathbb{E}_i$. De même, la projection de μ_i sur $\mathbb{E}_i^\perp$ est également μ_i, car $\mu_i \in \mathbb{E}_i^\perp$.

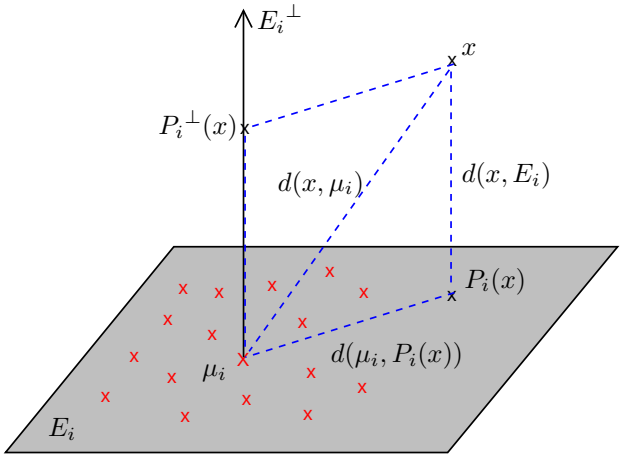


FIG. 3.1 – L'espace $\mathbb{E}_i$ caractéristique de la classe C_i, $\forall i = 1, ..., k$.

3.2 Construction de la règle de décision

Nous avons donc écrit la matrice de variance Δ_i comme composée de la matrice de variance dans $\mathbb{E}_i$ et de la matrice de variance dans $\mathbb{E}_i^{\perp}$. Ainsi, sur le modèle de l'Analyse Discriminante classique et fort de nos hypothèses, nous allons pouvoir construire une nouvelle règle de décision.

Théorème 3.4. *Avec les notations et hypothèses précédentes sur Δ_i, $\forall i = 1, ..., k$, la règle δ^* donne lieu à une nouvelle règle de décision δ^+qui consiste à affecter x à la classe C_{i^*} si :*

$$i^* = \underset{i=1,\dots,k}{\operatorname{argmin}} \left\{ \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i) \right\}.$$

Démonstration. Nous allons réécrire la règle δ^* avec les hypothèses précédentes. La règle δ^*affecte x à la classe C_{i^*} si :

$$i^* = \underset{i=1,\dots,k}{\operatorname{argmin}}\{K_i(x)\},$$

où $K_i(x) = -2\log(\pi_i p(x|C_i))$ est la fonction de *coût* introduite à la définition 2.1. Or,

$$-2\log(p(x|C_i)) = (x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) + \log(\det \Sigma_i) + p\log(2\pi), \tag{3.2}$$

où $p\log(2\pi)$ est une constante commune à toutes les classes. En remplaçant Σ_i^{-1} par sa valeur en fonction de Δ_i dans l'expression $(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i)$, on obtient :

$$(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) \quad = \quad (x - \mu_i)^t (Q_i \Delta_i Q_i^t)^{-1}(x - \mu_i).$$

Or $Q_i^t Q_i = Id$ et par conséquent :

$$\begin{aligned}(x - \mu_i)^t \Sigma_i^{-1}(x - \mu_i) &= (x - \mu_i)^t Q_i \Delta_i^{-1} Q_i^t (x - \mu_i), \\ &= \left[Q_i^t(x - \mu_i)\right]^t \Delta_i^{-1} \left[Q_i^t(x - \mu_i)\right].\end{aligned}$$

Étant donné la structure de la matrice Δ_i (voir (3.1)), on peut décomposer Δ_i de la façon suivante :

$$\Delta_i = a_i A_i + b_i B_i, \tag{3.3}$$

où A_i est la matrice diagonale contenant des 1 sur les d_i premières lignes et des 0 ailleurs et $B_i = Id - A_i$. En passant à l'inverse, on obtient :

$$\Delta_i^{-1} = \frac{1}{a_i} A_i + \frac{1}{b_i} B_i,$$

et

$$\begin{aligned}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) &= \frac{1}{a_i}\left[Q_i^t(x-\mu_i)\right]^t A_i\left[Q_i^t(x-\mu_i)\right] \\ &+ \frac{1}{b_i}\left[Q_i^t(x-\mu_i)\right]^t B_i\left[Q_i^t(x-\mu_i)\right].\end{aligned}$$

Or, $A_i = A_iA_i^t$ et $B_i = B_iB_i^t$. Donc,

$$\begin{aligned}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) &= \frac{1}{a_i}\left[Q_i^t(x-\mu_i)\right]^t A_iA_i^t\left[Q_i^t(x-\mu_i)\right] \\ &+ \frac{1}{b_i}\left[Q_i^t(x-\mu_i)\right]^t B_iB_i^t\left[Q_i^t(x-\mu_i)\right],\end{aligned}$$

et l'on obtient :

$$(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) = \frac{1}{a_i}\|(Q_iA_i)^t(x-\mu_i)\|^2 + \frac{1}{b_i}\|(Q_iB_i)^t(x-\mu_i)\|^2.$$

On remarque que $Q_iA_i = \tilde{Q}_i$ et $Q_iB_i = (Q_i - \tilde{Q}_i)$. Par conséquent, on peut écrire :

$$\begin{aligned}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) &= \frac{1}{a_i}\|\tilde{Q}_i{}^t(x-\mu_i)\|^2 + \frac{1}{b_i}\|(Q_i-\tilde{Q}_i)^t(x-\mu_i)\|^2, \\ &= \frac{1}{a_i}\|\tilde{Q}_i\tilde{Q}_i{}^t(x-\mu_i)\|^2 + \frac{1}{b_i}\|(Q_i-\tilde{Q}_i)(Q_i-\tilde{Q}_i)^t(x-\mu_i)\|^2.\end{aligned}$$

En utilisant la définition 3.3, on obtient que $\tilde{Q}_i\tilde{Q}_i{}^t(x-\mu_i) = P_i(x)-\mu_i$ et $(Q_i-\tilde{Q}_i)(Q_i-\tilde{Q}_i)^t(x-\mu_i) = P_i^\perp(x)-\mu_i$. La figure 3.1 permet de nous convaincre que $P_i^\perp(x)-\mu_i = x - P_i(x)$ et l'on obtient :

$$(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x-P_i(x)\|^2.$$

Par conséquent, la fonction de coût K_i de l'équation (3.2) s'écrit à présent :

$$K_i(x) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x-P_i(x)\|^2 + \log(\det\Sigma_i) - 2\log(\pi_i) + C^{te}. \tag{3.4}$$

Il ne nous reste plus qu'à calculer le déterminant de la matrice de variance Σ_i :

$$\det\Sigma_i = \det\Delta_i = (a_i)^{d_i}(b_i)^{(p-d_i)},$$

et par suite,

$$\log(\det\Sigma_i) = d_i\log(a_i) + (p-d_i)\log(b_i).$$

En remplaçant la valeur de $\det\Sigma_i$ dans l'équation (3.4) cela conduit à la nouvelle écriture de la fonction de coût K_i :

$$K_i(x) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x-P_i(x)\|^2 + d_i\log(a_i) + (p-d_i)\log(b_i) - 2\log(\pi_i) + C^{te}.$$

□

3.3 Retour à la probabilité *a posteriori*

Il peut être particulièrement intéressant de disposer de la probabilité *a posteriori* $p(C_i|x)$ que le point x appartienne à la classe C_i pour connaître l'incertitude de classification d'un point x.

Proposition 3.5. *La probabilité* a posteriori $p(C_i|x)$ *que le point* x *appartienne à la classe* C_i *est donnée par :*

$$p(C_i|x) = \frac{\exp\left(-\frac{1}{2}K_i(x)\right)}{\sum_{j=1}^{k}\exp\left(-\frac{1}{2}K_j(x)\right)},$$

où K_i *est la fonction de coût relative à la classe* C_i *:*

$$K_i(x) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + d_i\log(a_i) + (p-d_i)\log(b_i) - 2\log(\pi_i) + C^{te}.$$

Démonstration. Nous avons vu que la règle δ^+ ne fait pas directement appel à $p(C_i|x)$, mais à la fonction coût K_i relative à la classe C_i :

$$K_i(x) = -2\log(\pi_i p(x|C_i)).$$

Par conséquent et en appliquant la formule de Bayes, on obtient le résultat :

$$p(C_i|x) = \frac{\pi_i\, p(x|C_i)}{p(x)} = \frac{\exp\left(-\frac{1}{2}K_i(x)\right)}{\sum_{j=1}^{k}\exp\left(-\frac{1}{2}K_j(x)\right)}.$$

□

Remarque 3. La probabilité d'erreur de classification du point x est égale à $1 - p(C_{i^*}|x)$, où C_{i^*} est la classe à laquelle il a été affecté.

3.4 Reformulation de la règle δ^+

La règle δ^+ que nous avons énoncée dans ces pages peut être reformulée dans le but de faciliter son interprétation. Pour cela, nous avons besoin des notations suivantes. On pose :

$$\forall i = 1, ..., k, \begin{cases} a_i = \frac{\sigma_i^2}{\alpha_i}, \\ \\ b_i = \frac{\sigma_i^2}{(1-\alpha_i)}, \end{cases}$$

avec $\alpha_i \in]0,1[$ et $\sigma_i > 0$.

Corollaire 3.6. *Ces notations et hypothèses permettent de réécrire la règle* δ^+ *sous la forme suivante :*

$$x \in C_{i^*} \quad \text{si} \quad i^* = \underset{i=1,\ldots,k}{\operatorname{argmin}} \left\{ \frac{1}{\sigma_i^2}\left(\alpha_i\|\mu_i - P_i(x)\|^2 + (1-\alpha_i)\|x - P_i(x)\|^2\right) \right.$$
$$\left. +2p\log(\sigma_i) + d_i\log\left(\frac{1-\alpha_i}{\alpha_i}\right) - p\log(1-\alpha_i) - 2\log(\pi_i)\right\}.$$

Démonstration. Nous allons simplement effectuer le changement de variables $a_i = \frac{\sigma_i^2}{\alpha_i}$ et $b_i = \frac{\sigma_i^2}{(1-\alpha_i)}$ dans l'expression de K_i obtenue au théorème 3.4. Afin de simplifier les écritures, nous noterons $\phi_i(x) = \|\mu_i - P_i(x)\|^2$ et $\psi_i(x) = \|x - P_i(x)\|^2$. On peut alors écrire :

$$\begin{aligned} K_i(x) &= \frac{1}{a_i}\phi_i(x) + \frac{1}{b_i}\psi_i(x) + d_i\log(a_i) + (p-d_i)\log(b_i) - 2\log(\pi_i) + C^{te}, \\ &= \frac{\alpha_i}{\sigma_i^2}\phi_i(x) + \frac{(1-\alpha_i)}{\sigma_i^2}\psi_i(x) + d_i\log\left(\frac{\sigma_i^2}{\alpha_i}\right) + (p-d_i)\log\left(\frac{\sigma_i^2}{1-\alpha_i}\right) - 2\log(\pi_i) + C^{te}, \\ &= \frac{1}{\sigma_i^2}\left[\alpha_i\phi_i(x) + (1-\alpha_i)\psi_i(x)\right] + 2p\log(\sigma_i) + d_i\log\left(\frac{1-\alpha_i}{\alpha_i}\right) \\ &\quad -p\log(1-\alpha_i) - 2\log(\pi_i) + C^{te}, \end{aligned}$$

ce qui permet d'obtenir le résultat. □

Nb. de paramètres $\chi(k,p)$	$[Qd]$	$[Qd_i]$	$[Q_id]$	$[Q_id_i]$
Modèle $[ab]$ (cl. isométriques)	$\rho+\tau+3$ (2888, FE)	$\rho+\tilde{\tau}+k+2$ (2891, MI)	$\rho+k\tau+3$ (9998, FE)	$\rho+\bar{\tau}+k+2$ (10001, FE)
Modèle $[a_ib]$ (bruit commun)	$\rho+\tau+k+2$ (2891, MI)	$\rho+\tilde{\tau}+2k+1$ (2894, MI)	$\rho+k(\tau+1)+2$ (10001, FE)	$\rho+\bar{\tau}+2k+1$ (10004, FE)
Modèle $[ab_i]$	$\rho+\tau+k+2$ (2891, MI)	$\rho+\tilde{\tau}+2k+1$ (2894, MI)	$\rho+k(\tau+1)+2$ (10001, FE)	$\rho+\bar{\tau}+2k+1$ (10004, FE)
Modèle $[\alpha_i\sigma]$	$\rho+\tau+k+2$ (2891, MI)	$\rho+\tilde{\tau}+2k+1$ (2894, MI)	$\rho+k(\tau+1)+2$ (10001, MI)	$\rho+\bar{\tau}+2k+1$ (10004, MI)
Modèle $[\alpha\sigma_i]$ (cl. homothétiques)	$\rho+\tau+k+2$ (2891, MI)	$\rho+\tilde{\tau}+2k+1$ (2894, MI)	$\rho+k(\tau+1)+2$ (10001, MI)	$\rho+\bar{\tau}+2k+1$ (10004, MI)
Modèle $[a_ib_i]$	$\rho+\tau+2k+1$ (2894, MI)	$\rho+\tilde{\tau}+3k$ (2897, MI)	$\rho+k(\tau+2)+1$ (10004, FE)	**HDDA** $\rho+\bar{\tau}+3k$ (10007, FE)
Méthodes de référence	**LDAs** $\rho+1$ $\chi(4,128)=516$	**QDAs** $\rho+k$ $\chi(4,128)=519$	**LDA** $\rho+p(p+1)/2$ $\chi(4,128)=8771$	**QDA** $\rho+kp(p+1)/2$ $\chi(4,128)=33539$

TAB. 4.1 – Les différents cas particuliers de l'HDDA : $\chi(k,p)$ est le nombre de paramètres à estimer, où $\rho = kp+k-1$ est le nombre de paramètres nécessaires à l'estimation des moyennes et proportions et $\tau_i = d_i\,[p-(d_i-1)/2]$ est le nombre de paramètres nécessaires à l'estimation des d_i premières colonnes d'une matrice orthogonale. Nous noterons $\tilde{\tau} = \max_{i=1,\ldots,k}(\tau_i)$, $\bar{\tau} = \sum_{i=1}^{k}\tau_i$ et τ le nombre de paramètres nécessaires à l'estimation des d premières colonnes d'une matrice orthogonale quand les dimensions d_i sont communes et égales à d. Il est indiqué entre parenthèses la valeur de $\chi(4,128)$ avec $\forall i,\ d_i = 20$ et si les estimateurs du modèle ont une forme explicite (FE) ou s'ils sont déterminés par une méthode itérative (MI).

4 Règles particulières de l'HDDA

La méthode que nous proposons dans ce rapport peut engendrer des règles de décision interprétables de façon simple pour des valeurs particulières des différents paramètres. Ces règles particulières peuvent être vues comme autant de régularisations possibles de l'HDDA dans le sens où elles font des hypothèses supplémentaires sur les paramètres et peuvent ainsi permettre une meilleure estimation de ces derniers. Le tableau 4.1 résume les différents cas particuliers de l'HDDA existants et indique en particulier le nombres de paramètres à estimer pour ces modèles. On remarque notamment que le nombre de paramètres à estimer dépend linéairement de la dimension de l'espace initial et non pas quadratiquement comme pour QDA et LDA.
Avant d'expliciter les règles découlant de la règle δ^+ dans le cas où les classes sont isométriques et homothétiques, nous présenterons les liens qui existent entre l'Analyse Discriminante de Haute Dimension, l'Analyse Discriminante Linéaire et l'Analyse Discriminante à Décomposition Spectrale. La figure 4.1 présente les liens qui existent entre les différentes méthodes d'Analyse Discriminante.

4.1 Liens avec l'Analyse Discriminante classique

Nous allons tout d'abord expliciter les liens qui existent entre l'Analyse Discriminante de haute dimension et l'Analyse Discriminante classique.

Proposition 4.1. *La règle δ^+ est équivalente aux règles de l'Analyse Discriminante classique, présentées au paragraphe 2.3, dans les cas suivants :*

(i) si $\forall i = 1, ..., k$, $\alpha_i = \frac{1}{2}$: la règle δ^+ est équivalente à la règle quadratique avec un modèle de classes homothétique et sphériques (QDAs),

(ii) si de plus $\forall i = 1, ..., k$, $\sigma_i = \sigma$: la règle δ^+ est équivalente à la règle linéaire avec un modèle de classes isométriques et sphériques (LDAs),

(iii) si de plus $\forall i = 1, ..., k$, $\pi_i = \pi_*$: la règle δ^+ est équivalente à la règle géométrique (LDAgéo).

Démonstration. En remplaçant, dans la formulation de K_i obtenue au corollaire 3.6, les paramètres α_i, σ_i et π_i par leurs nouvelles valeurs énoncées ci-dessus, on obtient :

$$\begin{aligned} K_i(x) &= \frac{1}{2\sigma_i^2}\left(\|\mu_i - P_i(x)\|^2 + \|x - P_i(x)\|^2\right) + 2p\log(2\sigma_i) \\ &\quad + d_i\log(1) - p\log(\frac{1}{2}) - 2\log(\pi_i) + C^{te}, \end{aligned}$$

ce qui est égal, à un facteur multiplicatif près, à :

$$K_i(x) = \frac{1}{\sigma_i^2}\left(\|\mu_i - P_i(x)\|^2 + \|x - P_i(x)\|^2\right) + 2p\log(\sigma_i) - 2\log(\pi_i) + C^{te}.$$

Le théorème de Pythagore nous permet de retrouver l'équation (2.6) et donc d'obtenir le résultat (i) :

$$K_i(x) = \frac{1}{\sigma_i^2}\|x - \mu_i\|^2 + 2p\log(\sigma_i) - 2\log(\pi_i) + C^{te}.$$

Si l'on ajoute à présent la contrainte $\forall i = 1, ..., k$, $\sigma_i = \sigma$, alors on peut écrire :

$$K_i(x) = \frac{1}{\sigma^2}\|x - \mu_i\|^2 - 2\log(\pi_i) + C^{te},$$

ce qui nous permet de retrouver l'équation (2.9) et donc d'obtenir le résultat (ii). Enfin, si l'on ajoute à présent la contrainte $\forall i = 1, ..., k$, $\pi_i = \pi_*$, alors on peut écrire :

$$K_i(x) = \frac{1}{\sigma^2}\|x - \mu_i\|^2 + C^{te},$$

ce qui donne l'équation (2.10) et donc le résultat (iii). □

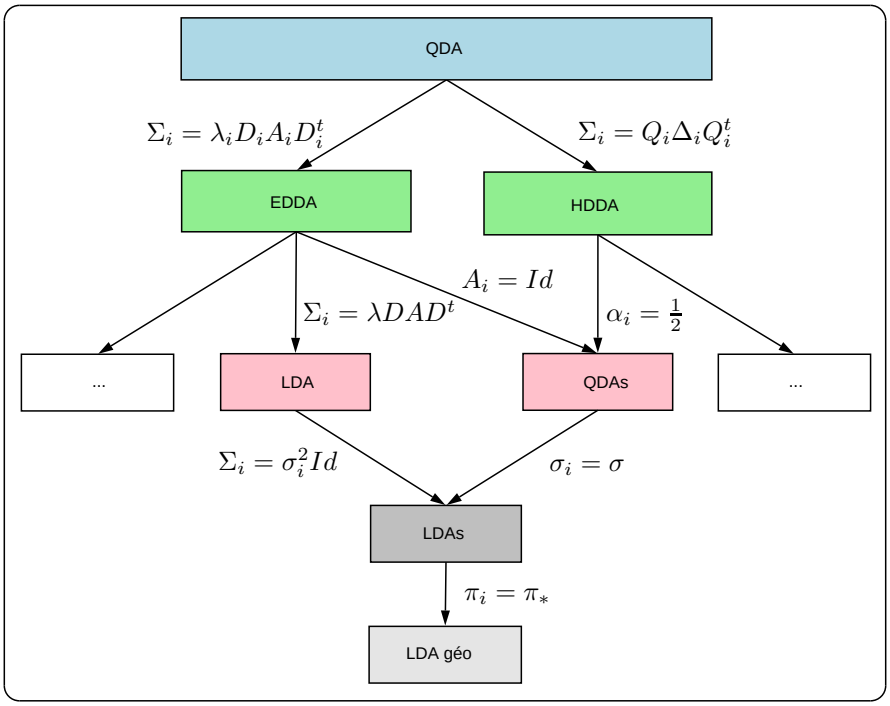


FIG. 4.1 – Liens entre les différentes méthodes d'Analyse Discriminante.

4.2 Liens avec l'Analyse Discriminante à Décomposition Spectrale

L'HDDA peut être également mise en relation avec l'EDDA présentée au paragraphe 2.4. En effet, ces deux méthodes ont en commun le fait de reparamétriser les matrices de covariance des classes. Il nous est donc aisé d'écrire notre paramétrisation des matrices de covariance avec les conventions de l'EDDA. On a, $\forall i = 1, ..., k,$:

$$\Sigma_i = Q_i \Delta_i Q_i^t,$$

or, d'après l'équation (3.3),

$$\Delta_i = a_i A_i + b_i B_i,$$

ce qui permet d'obtenir l'écriture suivante de la matrice de covariance de la classe C_i :

$$\Sigma_i = a_i Q_i A_i Q_i^t + b_i Q_i B_i Q_i^t.$$

Ainsi, nous avons reparamétrisé les matrices de covariance des classes par a_i, b_i, A_i, B_i et Q_i. Chacun de ces paramètres permet de contrôler une des caractéristiques de la distribution de la classe C_i :

(i) a_i et b_i permettent de contrôler respectivement le volume de la classe dans l'espace $\mathbb{E}_i$ et dans l'espace $\mathbb{E}_i^\perp$,

(ii) d_i permet de contrôler la forme de la classe dans l'espace $\mathbb{E}_i$ et dans l'espace $\mathbb{E}_i^\perp$ *via* A_i et B_i,

(iii) Q_i permet de contrôler l'orientation générale de la classe.

Ainsi le modèle général de l'HDDA peut être identifié par la notation $[a_i b_i Q_i d_i]$ ou de façon équivalente $[\alpha_i \sigma_i Q_i d_i]$. Avec ces conventions et de la même façon que l'EDDA, on peut identifier un certain nombre de modèles particuliers régularisant le modèle général de l'HDDA. Le tableau 4.1 liste ces différents cas particuliers et on peut ainsi dénombrer 23 modèles différents issus de l'HDDA. Dans les paragraphes suivants, nous avons choisit de nous intéresser plus particulièrement à 2 d'entre eux : $[\alpha \sigma Q_i d]$ et $[\alpha \sigma_i Q_i d]$. Le tableau 4.2 présente les différentes variantes existantes pour ces deux modèles.

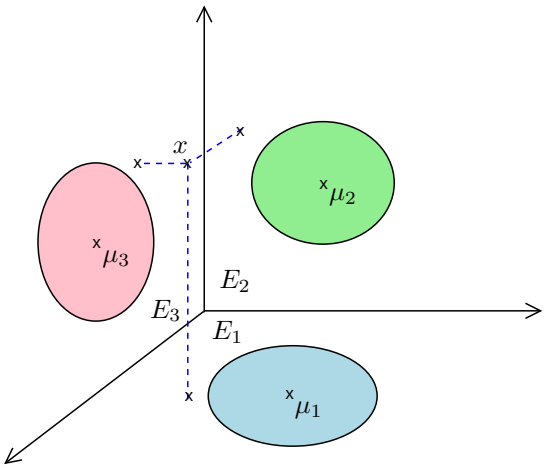


FIG. 4.2 – Modèle $[abQ_id]$: les classes sont isométriques (*i.e.* $\forall i,\ \sigma_i = \sigma$) et la règle de décision ne tient donc compte que des distances $d(x, E_i)$ et $d(P_i(x), \mu_i)$.

4.3 Règle isométrique de décision : modèle $[\alpha\sigma Q_i d]$

Afin d'expliciter la règle de décision δ^+ dans le cas de classes isométriques, nous allons nous placer dans le cas où les paramètres α_i, σ_i, π_i et d_i sont respectivement égaux, *i.e.* :

$$\forall i = 1, ..., k, \begin{cases} \alpha_i = \alpha, \\ \sigma_i = \sigma, \\ d_i = d, \\ \pi_i = \pi_*. \end{cases} \tag{4.1}$$

La figure 4.2 illustre le cas présent où les classes sont isométriques.

Proposition 4.2. *Sous les hypothèses (4.1), nous nous plaçons dans le cadre de classes isométriques et la règle de décision δ^+ s'écrit alors :*

$$x \in C_{i^*} \quad \text{si} \quad i^* = \underset{i=1,...,k}{\operatorname{argmin}} \left\{ \alpha \|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 \right\}. \tag{4.2}$$

Démonstration. En remplaçant, dans la formulation de K_i obtenue au corollaire 3.6, les paramètres α_i, σ_i et π_i par leurs nouvelles valeurs énoncées ci-dessus, on obtient :

$$\begin{aligned} K_i(x) &= \frac{1}{\sigma^2}\left(\alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2\right) + 2p\log(\sigma) \\ &\quad + d\log\left(\frac{1-\alpha}{\alpha}\right) - p\log(1-\alpha) - 2\log(\pi_*) + C^{te}, \end{aligned}$$

ce qui est équivalent, à une constante multiplicative près et en regroupant les quantités indépendantes de la classe, à :

$$K_i(x) = \alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 + C^{te}.$$

□

Cas où $\alpha = 0$ La règle de décision (4.2) consiste à affecter le point x à la classe C_{i^*} si $\forall i = 1, ..., k$, $d(x, \mathbb{E}_{i^*}) \leq d(x, \mathbb{E}_i)$. D'un point de vue géométrique on affecte x à C_{i^*} si il est plus proche du sous espace propre de cette classe que des autres (voir Fig. 4.2).

α_i ***égaux***	σ_i ***égaux***	σ_i ***libres***
d_i ***et*** π_i ***égaux***	Modèle $[abQ_id]$: - Classes isométriques - $\delta^+ = \boxed{1}$	Modèle $[\alpha\sigma_iQ_id]$: - Classes homothétiques - $\delta^+ = \boxed{1} + \boxed{2}$
Options :		
d_i *libres :*	$\delta^+ = \boxed{1} + \boxed{3}$	$\delta^+ = \boxed{1} + \boxed{2} + \boxed{3}$
π_i *libres :*	$\delta^+ = \boxed{1} + \boxed{5}$	$\delta^+ = \boxed{1} + \boxed{2} + \boxed{5}$
d_i *et* π_i *libres :*	$\delta^+ = \boxed{1} + \boxed{3} + \boxed{5}$	$\delta^+ = \boxed{1} + \boxed{2} + \boxed{3} + \boxed{5}$

α_i***,*** d_i ***et*** π_i ***libres***

$\downarrow$

$HDDA$:
- $\delta^+ = \boxed{1} + \boxed{2} + \boxed{3} + \boxed{4} + \boxed{5}$

On rappelle que la règle de décision δ^+ consiste à affecter x à la classe C_{i^*} si $i^* = \text{argmin}_{i=1,\ldots,k}\{K_i(x)\}$, où :

$$K_i(x) = \underbrace{\frac{1}{\sigma_i^2}\left(\alpha_i\|\mu_i - P_i(x)\|^2 + (1-\alpha_i)\|x - P_i(x)\|^2\right)}_{\boxed{1}} + \underbrace{2p\log(\sigma_i)}_{\boxed{2}}$$
$$+ \underbrace{d_i \log\left(\frac{1-\alpha_i}{\alpha_i}\right)}_{\boxed{3}} - \underbrace{p\log(1-\alpha_i)}_{\boxed{4}} - \underbrace{2\log(\pi_i)}_{\boxed{5}} + C^{te}.$$

TAB. 4.2 – Variantes des modèles $[abQ_id]$ et $[\alpha\sigma_iQ_id]$.

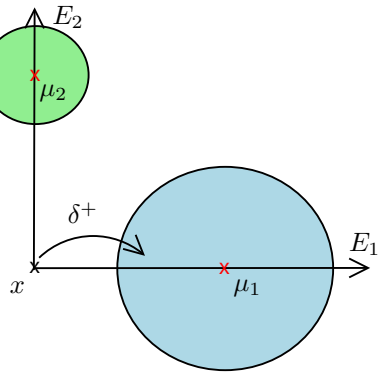


FIG. 4.3 – Modèle $[\alpha\sigma_i Q_i d]$: le point x, étant équidistant de μ_1 et μ_2, sera affecté à la classe C_1 car celle-ci est plus « attractive » du fait de sa variance plus grande.

Cas où $\alpha = 1$ La règle de décision (4.2) consiste à affecter le point x à la classe C_{i^*} si $\forall i = 1, ..., k$, $d(P_{i^*}(x), \mu_{i^*}) \leq d(P_i(x), \mu_i)$. Cela signifie que l'on affecte x à C_{i^*} si sa projection dans le sous espace propre de cette classe est plus proche du barycentre de cette classe que sa projection dans les autres espaces propres l'est des autres barycentres (voir Fig. 4.2).

Cas où $0 < \alpha < 1$ L'hypothèse de normalité des classes implique que les deux valeurs précédentes de α ne conduiront pas à des règles optimales de décision. L'estimation de α est faite au paragraphe 5. Ainsi, la règle de décision affectera x à la classe réalisant le meilleur compromis entre les 2 cas précédents.

4.4 Règle homothétique de décision : modèle $[\alpha\sigma_i Q_i d]$

Cette règle diffère de la règle isométrique présentée précédemment du fait qu'elle relaxe la contrainte d'égalité imposée au paramètre σ_i. Nous sommes alors dans le cas où uniquement les paramètres α_i et d_i sont respectivement égaux.

Proposition 4.3. *Sous ces hypothèses, nous nous plaçons dans le cadre de classes homothétiques et la règle de décision δ^+ s'écrit :*

$$x \in C_{i^*} \quad \text{si} \quad i^* = \underset{i=1,\ldots,k}{\operatorname{argmin}} \left\{ \frac{1}{\sigma_i^2} \left(\alpha \|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 \right) + 2p \log(\sigma_i) \right\}.$$

Démonstration. En remplaçant, dans la formulation de K_i obtenue au corollaire 3.6, les paramètres α_i, σ_i et π_i par leurs valeurs dans ce cas, on obtient :

$$\begin{aligned} K_i(x) &= \frac{1}{\sigma_i^2} \left(\alpha \|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 \right) + 2p \log(\sigma_i) \\ &\quad + d \log\left(\frac{1-\alpha}{\alpha} \right) - p \log(1-\alpha) - 2 \log(\pi_*) + C^{te}, \\ &= \frac{1}{\sigma_i^2} \left(\alpha \|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 \right) + 2p \log(\sigma_i) + C^{te}. \end{aligned}$$

□

Ainsi, la règle de décision va favoriser les classes de grande variance. En effet, si un point x est à la même « distance » de deux classes, il est naturel qu'il soit affecté à la classe de plus grande variance. Ce cas de figure est illustré par la figure 4.3.

4.5 Relaxe des contraintes d'égalité portant sur les d_i et π_i

Règles particulières avec d_i libres Les règles précédentes font l'hypothèse que les dimensions intrinsèques des espaces propres des classes sont égales. Toutefois, cette hypothèse peut se révéler trop restrictive si les dimensions intrinsèques sont significativement différentes. Si l'on relaxe cette hypothèse, la règle δ^+ prévoit un terme de pénalisation en fonction de la dimension intrinsèque des espaces propres et de la valeur de α. En effet, la fonction de coût K_i contient alors en plus la quantité Θ_i suivante :

$$\Theta_i = d_i \log\left(\frac{1-\alpha}{\alpha}\right).$$

Ainsi, si $\alpha < \frac{1}{2}$, la règle de décision donne un poids plus important à la distance entre x et l'espace $\mathbb{E}_i$ et il faut donc pénaliser les espaces de grande dimension. En effet, un point quelconque de $\mathbb{R}^p$ est généralement plus près d'un espace de grande dimension que d'un espace de petite dimension. Au contraire, si $\alpha > \frac{1}{2}$, la règle de décision donne un poids plus important à la distance entre la projection de x sur $\mathbb{E}_i$ et le barycentre de la classe et il faut donc pénaliser les espaces de petite dimension.

Règles particulières avec π_i libres Les règles précédentes sont pénalisées, comme leur équivalent de l'Analyse Discriminante Linéaire, dans le cas où l'hypothèse d'égalité des probabilités *a priori* π_i est fausse. Pour pallier cette limitation on peut choisir de ne pas prendre $\forall i = 1, ..., k$, $\pi_i = \frac{1}{k}$. En effet, la fonction de coût K_i contient alors en plus la quantité Ξ_i suivante :

$$\Xi_i = -2\log(\pi_i).$$

Ainsi, la règle de décision favorisera les classes dont la probabilité *a priori* π_i est grande (*i.e.* proche de 1).

4.6 Règles particulières avec $Q_i = Q$

Si l'on considère des données pour lesquelles il est raisonnable de penser que l'orientation générale des classes est commune, alors on peut faire l'hypothèse que les matrices Q_i sont égales. Nous nous focalisons ici sur trois des modèles :

Modèle $[a_i b_i Q d_i]$ L'orientation de la classe C_i étant contrôlée par la matrice Q_i, si l'orientation générale des classes est commune alors il convient de chercher une matrice Q commune telle que $\forall i = 1, ..., k$, $\Delta_i = Q^t \Sigma_i Q$. Ce modèle a été baptisé *common principal component* par Flury [10] et l'estimation de la matrice Q doit être faite par une procédure itérative (voir paragraphe 5).

Modèle $[abQd]$ Ce cas particulier diffère du modèle $[\alpha\sigma Q_i d_i]$, *i.e.* de la règle isométrique, du fait que tous les paramètres sont supposés communs. Nous considérons alors que les classes sont de même forme, de même orientation et vivent dans des sous-espaces de même dimension intrinsèque d.

Proposition 4.4. *Sous ces hypothèses, la règle de décision δ^+ s'écrit alors :*

$$x \in C_{i^*} \quad \text{si} \quad i^* = \underset{i=1,\ldots,k}{\operatorname{argmin}} \left\{ \frac{1}{a}\|\mu_i - P_i(x)\|^2 + \frac{1}{b}\|x - P_i(x)\|^2 \right\},$$

où $P_i(x) = \tilde{Q}\tilde{Q}^t(x - \mu_i) + \mu_i$.

Démonstration. Ce résultat s'obtient facilement en remplaçant dans l'expression de δ^+ obtenue au théorème 3.4 les paramètres a_i, b_i et π_i par leurs valeurs dans ce cas. L'hypothèse d'égalité des Q_i se traduit par la modification de l'opérateur de projection P_i qui devient $P_i(x) = \tilde{Q}\tilde{Q}^t(x - \mu_i) + \mu_i$. Les quantités indépendantes de la classe peuvent être omises car elles n'interviennent pas dans la décision. □

Modèle $[\alpha\sigma_i Qd]$ Ce cas particulier diffère du modèle $[\alpha\sigma_i Q_i d_i]$, *i.e.* de la règle homothétique, du fait que les classes sont supposées avoir la même orientation et vivre dans des sous-espaces de même dimension intrinsèque d.

Proposition 4.5. *Sous ces hypothèses, la règle de décision δ^+ s'écrit alors :*

$$x \in C_{i^*} \quad \text{si} \quad i^* = \underset{i=1,\ldots,k}{\operatorname{argmin}} \left\{ \frac{1}{\sigma_i^2} \left(\alpha \|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2 \right) + 2p \log \sigma_i \right\},$$

où $P_i(x) = \tilde{Q}\tilde{Q}^t(x - \mu_i) + \mu_i$.

5 Estimation des paramètres

La règle de décision δ^+ que nous avons présentée précédemment requiert l'estimation de certains paramètres. Dans ce chapitre, outre la détermination de la dimension intrinsèque d_i de chaque classe qui est un problème difficile et qui sera traitée à la fin de ce chapitre, il nous faudra également estimer les paramètres apparaissant dans la règle δ^+ de l'HDDA ainsi que dans ses variantes. Ce chapitre est organisé de la façon suivante : nous présenterons tout d'abord les estimateurs communs à tous les modèles puis ceux de l'HDDA et enfin des cas particuliers de l'HDDA.

5.1 Estimateurs communs

Dans ce rapport, nous avons choisi d'estimer les probabilités *a priori* des groupes par leur proportion :

$$\forall i = 1, ..., k,\ \hat{\pi_i} = \frac{n_i}{n},$$

où $n_i = Card(C_i)$. De même, les moyennes et les matrices de covariance des classes sont estimées classiquement par :

$$\hat{\mu_i} = \bar{x_i} = \frac{1}{n_i} \sum_{x_j \in C_i} x_j,$$

$$\hat{\Sigma_i} = \frac{1}{n_i} \sum_{x_j \in C_i} (x_j - \hat{\mu_i})^t (x_j - \hat{\mu_i}).$$

D'autre part, certains des estimateurs présentés dans la suite de ce paragraphe s'expriment en fonction des $(p - d_i)$ plus petites valeurs propres de $\hat{\Sigma_i}$. Afin de minimiser le nombre de paramètres à estimer, nous ne déterminerons pas explicitement ces valeurs propres ni les vecteurs propres associés. Il est donc uniquement nécessaire d'estimer les d_i premières colonnes de la matrice Q_i, ce qui représente une économie importante dans le nombre de paramètres à estimer (voir tableau 4.1). La quantité $\sum_{l=d_i+1}^{p} \lambda_{il}$ sera donc calculée grâce à la relation suivante qui ne fait intervenir que les d_i plus grandes valeurs propres de $\hat{\Sigma_i}$:

$$\sum_{l=d_i+1}^{p} \lambda_{il} = \mathrm{tr}(\hat{\Sigma_i}) - \sum_{l=1}^{d_i} \lambda_{il}.$$

5.2 Estimateurs de l'HDDA

Dans le but de faciliter la démonstration des résultats suivants, nous allons établir l'écriture de la log-vraisemblance de la classe C_i (issue de [10, eq. (2.5)]).

Lemme 5.1. *La log-vraisemblance de la classe C_i, $\forall i = 1, ..., k$ vérifie la relation suivante :*

$$-2\log(L_i(x_j \in C_i, \mu_i, \Sigma_i)) = n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma_i} q_{il} \right) + C^{te},$$

où δ_{il} est le $l^{\text{ème}}$ terme de la matrice diagonale Δ_i et q_{il} est la $l^{\text{ème}}$ colonne de la matrice Q_i.

Démonstration. Avec les hypothèses faites dans les paragraphes précédents, la vraisemblance du modèle de la classe C_i vaut, $\forall i = 1, ..., k$:

$$L_i(x_j \in C_i, \mu_i, \Sigma_i) = \prod_{x_j \in C_i} \frac{1}{(2\pi)^{p/2}(\det \Sigma_i)^{1/2}} \exp\left(-\frac{1}{2}(x_j - \mu_i)^t \Sigma_i^{-1}(x_j - \mu_i)\right),$$

et par conséquent :

$$-2\log(L_i(x_j \in C_i, \mu_i, \Sigma_i)) = \sum_{x_j \in C_i} \left(\log(\det \Sigma_i) + (x_j - \mu_i)^t \Sigma_i^{-1}(x_j - \mu_i)\right) + C^{te}.$$

Or, on a la relation $\Sigma_i = Q_i \Delta_i Q_i^t$, où Δ_i est diagonale composée des termes diagonaux δ_{il} et où Q_i est composée des colonnes q_{il}, $l = 1, ..., p$. Cela nous permet d'écrire :

$$\begin{aligned}
-2\log(L_i) - C^{te} &= n_i \log(\prod_{l=1}^{p} \delta_{il}) + \sum_{x_j \in C_i} \operatorname{tr}\left((x_j - \mu_i)^t \Sigma_i^{-1}(x_j - \mu_i)\right), \\
&= n_i \sum_{l=1}^{p} \log(\delta_{il}) + \sum_{x_j \in C_i} \operatorname{tr}\left(\Sigma_i^{-1}(x_j - \mu_i)(x_j - \mu_i)^t\right), \\
&= n_i \sum_{l=1}^{p} \log(\delta_{il}) + \operatorname{tr}\left(\Sigma_i^{-1} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^t\right),
\end{aligned}$$

ce qui donne l'expression suivante :

$$-2\log(L_i) - C^{te} = n_i \sum_{l=1}^{p} \log(\delta_{il}) + n_i \operatorname{tr}\left(\Sigma_i^{-1} \hat{\Sigma}_i\right).$$

En remplaçant Σ_i^{-1} par sa valeur en fonction de Δ_i, on obtient :

$$\begin{aligned}
-2\log(L_i) - C^{te} &= n_i \sum_{l=1}^{p} \log(\delta_{il}) + n_i \operatorname{tr}\left(Q_i \Delta_i^{-1} Q_i^t \hat{\Sigma}_i\right), \\
&= n_i \sum_{l=1}^{p} \log(\delta_{il}) + n_i \operatorname{tr}\left(\Delta_i^{-1} Q_i^t \hat{\Sigma}_i Q_i\right), \\
&= n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il}\right).
\end{aligned}$$

□

Nous avons choisi d'estimer la matrice Q_i ainsi que les paramètres a_i et b_i au sens du maximum de vraisemblance sur l'ensemble d'apprentissage A. Ces estimations concernent les modèles $[a_i b_i Q_i d_i]$ et $[a_i b_i Q_i d]$.

Proposition 5.2. *Les estimateurs au sens du maximum de vraisemblance de la matrice Q_i et des paramètres a_i et b_i de la classe C_i existent et sont uniques, $\forall i = 1, ..., k$:*

(i) Q_i est estimée par la matrice $\hat{Q}_i$ dont les d_i premières colonnes sont les vecteurs propres associés aux d_i plus grandes valeurs propres de $\hat{\Sigma}_i$ et les $(p - d_i)$ dernières colonnes sont les vecteurs propres associés aux $(p - d_i)$ plus petites valeurs propres de $\hat{\Sigma}_i$,

(ii) a_i est estimé par la moyenne des d_i plus grandes valeurs propres de $\hat{\Sigma}_i$:

$$\hat{a_i} = \frac{\sum_{l=1}^{d_i} \lambda_{il}}{d_i},$$

(iii) et b_i est estimé par la moyenne des $(p-d_i)$ plus petites valeurs propres de $\hat{\Sigma}_i$:

$$\hat{b_i} = \frac{\sum_{l=d_i+1}^{p} \lambda_{il}}{(p-d_i)},$$

où λ_{il} est la $l^{\text{ème}}$ plus grande valeur propre de $\hat{\Sigma}_i$.

Démonstration. Le lemme 5.1 nous permet d'écrire :

$$-2\log(L_i(x_j \in C_i, \mu_i, \Sigma_i)) = n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il} \right) + C^{te},$$

avec $\delta_{il} = a_i$ si $l \leq d_i$ et $\delta_{il} = b_i$ sinon. On souhaite minimiser cette quantité sous la contrainte $q_{il}^t q_{il} = 1$, ce qui revient à minimiser la fonction de Lagrange suivante :

$$\mathcal{L}_i = n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il} \right) - \sum_{l=1}^{p} \theta_{il}(q_{il}^t q_{il} - 1),$$

où les θ_{il} sont les multiplicateurs de Lagrange. La dérivée partielle de $\mathcal{L}_i$ par rapport à a_i vaut :

$$\frac{\partial \mathcal{L}_i}{\partial a_i} = n_i \sum_{l=1}^{p} \frac{\partial}{\partial \delta_{il}} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il} \right) \frac{\partial \delta_{il}}{\partial a_i},$$

avec $\frac{\partial \delta_{il}}{\partial a_i} = 1$ si $l \leq d_i$ et 0 sinon. On obtient :

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial a_i} &= n_i \sum_{l=1}^{d_i} \left(\frac{1}{a_i} - \frac{1}{a_i^2} q_{il}^t \hat{\Sigma}_i q_{il} \right), \\ &= \frac{n_i d_i}{a_i} - \frac{n_i}{a_i^2} \sum_{l=1}^{d_i} q_{il}^t \hat{\Sigma}_i q_{il}. \end{aligned}$$

La condition $\frac{\partial \mathcal{L}_i}{\partial a_i} = 0$ implique que :

$$\hat{a_i} = \frac{1}{d_i} \sum_{l=1}^{d_i} q_{il}^t \hat{\Sigma}_i q_{il}. \tag{5.1}$$

De même, la dérivée partielle de $\mathcal{L}_i$ par rapport à b_i vaut :

$$\frac{\partial \mathcal{L}_i}{\partial b_i} = n_i \sum_{l=1}^{p} \frac{\partial}{\partial \delta_{il}} \left(\log \delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il} \right) \frac{\partial \delta_{il}}{\partial b_i},$$

avec $\frac{\partial \delta_{il}}{\partial b_i} = 1$ si $l \geq d_i + 1$ et 0 sinon. Par conséquent, on a :

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial b_i} &= n_i \sum_{l=d_i+1}^{p} \left(\frac{1}{b_i} - \frac{1}{b_{il}^2} q_{il}^t \hat{\Sigma}_i q_{il} \right), \\ &= \frac{n_i(p-d_i)}{b_i} - \frac{n_i}{b_i^2} \sum_{l=d_i+1}^{p} q_{il}^t \hat{\Sigma}_i q_{il}. \end{aligned}$$

La condition $\frac{\partial \mathcal{L}_i}{\partial b_i} = 0$ implique que :

$$\hat{b_i} = \frac{1}{(p-d_i)} \sum_{l=d_i+1}^{p} q_{il}^t \hat{\Sigma}_i q_{il}. \tag{5.2}$$

Enfin, le gradient de $\mathcal{L}_i$ par rapport à q_{il} vaut :

$$\nabla_{q_{il}} \mathcal{L}_i = 2\frac{n_i}{\delta_{il}} \hat{\Sigma}_i q_{il} - 2\theta_{il} q_{il},$$

et en multipliant cette quantité à gauche par q_{il}^t, on a :

$$\begin{aligned} q_{il}^t \nabla_{q_{il}} \mathcal{L}_i = 0 \quad &\Leftrightarrow \quad 2\frac{n_i}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il} - 2\theta_{il} = 0, \\ &\Leftrightarrow \quad \theta_{il} = \frac{n_i}{\delta_{il}} q_{il}^t \hat{\Sigma}_i q_{il}, \end{aligned}$$

et par conséquent :

$$\hat{\Sigma}_i q_{il} = \frac{\theta_{il}\delta_{il}}{n_i} q_{il},$$

ce qui signifie que q_{il} est le vecteur propre de $\hat{\Sigma}_i$ associé à la valeur propre $\lambda_{il} = \frac{\theta_{il}\delta_{il}}{n_i}$. En reportant dans les expressions (5.1) et (5.2), cela permet d'établir les résultats (i) et (ii). Les q_{il} étant vecteurs propres de $\hat{\Sigma}_i$ qui est une matrice symétrique, cela implique que $q_{il}^t q_{ih} = 0$ si $h \neq l$ et que $q_{il}^t q_{il} = 1$. Il ne nous reste plus qu'à trouver l'ordre des vecteurs propres de $\hat{\Sigma}_i$ dans Q_i. On souhaite minimiser la quantité suivante à l'*optimum* :

$$-2\log L_i = n_i(d_i \log \hat{a_i} + (p - d_i)\log \hat{b_i}),$$

avec $\hat{a_i} > \hat{b_i}$ et $d_i\hat{a_i} + (p - d_i)\hat{b_i} = \sum_{l=1}^{p} \lambda_{il} = s_i$ qui est la trace de $\hat{\Sigma}_i$ et donc $\hat{a_i} > \frac{s_i}{p}$.

$$\frac{-2\partial \log L_i}{\partial \hat{a_i}} = \frac{d_i}{\hat{a_i}} + \frac{p - d_i}{s_i - d_i\hat{a_i}} < 0,$$

et donc, il faut choisir $\hat{a_i}$ le plus grand possible pour minimiser $-2\log L_i$. Cela ne peut être réalisé qu'en choisissant les vecteurs propres associés aux d_i plus grandes valeurs propres de $\hat{\Sigma}_i$ pour remplir les d_i premières colonnes de Q_i et, par conséquent, en choisissant les vecteurs propres associés aux $(p - d_i)$ plus petites valeurs propres de $\hat{\Sigma}_i$ pour remplir les $(p - d_i)$ dernières colonnes de Q_i. □

La proposition 5.2 nous permet de déduire les estimateurs des paramètres α_i et σ_i^2 dont nous aurons besoin pour les règles particulières :

Corollaire 5.3. *Les estimateurs au sens du maximum de vraisemblance des paramètres α_i et σ_i existent et sont uniques :*

$$\hat{\alpha_i} = \frac{\hat{b_i}}{\hat{a_i} + \hat{b_i}}, \tag{5.3}$$

$$\hat{\sigma_i^2} = \frac{\hat{a_i}\hat{b_i}}{\hat{a_i} + \hat{b_i}}. \tag{5.4}$$

5.3 Estimateurs des règles particulières à Q_i libres

Les règles particulières, énoncées au chapitre précédent, requièrent également l'estimation de certains paramètres. Nous présentons dans ce paragraphe les estimateurs des règles particulières ayant un modèle à Q_i libres.

Estimation de a : modèles $[ab_iQ_id_i]$ et $[ab_iQ_id]$

Proposition 5.4. *L'estimateur au sens du maximum de vraisemblance du paramètre a existe et est unique :*

$$\hat{a} = \frac{\sum_{i=1}^{k} n_i \sum_{l=1}^{d_i} \lambda_{il}}{\sum_{i=1}^{k} n_i d_i}.$$

Démonstration. La log-vraisemblance du modèle vérifie la relation suivante :

$$-2\log(L) = -2\sum_{i=1}^{k}\log(L_i),$$

ce qui, grâce au lemme 5.1 et au résultat (i) de la proposition 5.2, est égal à :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p}\left(\log\delta_{il} + \frac{1}{\delta_{il}}\lambda_{il}\right) + C^{te},$$

avec $\delta_{il} = a$ si $l \leq d_i$ et b_i sinon. Par conséquent, on peut écrire :

$$\begin{aligned}
-2\frac{\partial}{\partial a}\log(L) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{k} n_i \sum_{l=1}^{d_i}\left(\frac{1}{a} - \frac{1}{a^2}\lambda_{il}\right) = 0, \\
&\Leftrightarrow \quad \sum_{i=1}^{k} n_i d_i = \sum_{i=1}^{k} n_i \sum_{l=1}^{d_i}\frac{\lambda_{il}}{a}, \\
&\Leftrightarrow \quad a = \frac{\sum_{i=1}^{k} n_i \sum_{l=1}^{d_i}\lambda_{il}}{\sum_{i=1}^{k} n_i d_i},
\end{aligned}$$

ce qui permet de conclure. □

Estimation de b : modèles $[a_i b Q_i d_i]$ et $[a_i b Q_i d]$

Proposition 5.5. *L'estimateur au sens du maximum de vraisemblance du paramètre b existe et est unique :*

$$\hat{b} = \frac{\sum_{i=1}^{k} n_i \sum_{l=d_i+1}^{p}\lambda_{il}}{\sum_{i=1}^{k} n_i(p-d_i)}.$$

Démonstration. La log-vraisemblance du modèle vérifie la relation suivante :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p}\left(\log\delta_{il} + \frac{1}{\delta_{il}}\lambda_{il}\right) + C^{te},$$

avec $\delta_{il} = b$ si $l \geq d_i + 1$ et a_i sinon. Par conséquent, on peut écrire :

$$\begin{aligned}
-2\frac{\partial}{\partial b}\log(L) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{k} n_i \sum_{l=d_i+1}^{p}\left(\frac{1}{b} - \frac{1}{b^2}\lambda_{il}\right) = 0, \\
&\Leftrightarrow \quad \sum_{i=1}^{k} n_i(p-d_i) = \sum_{i=1}^{k} n_i \sum_{l=d_i+1}^{p}\frac{\lambda_{il}}{b},
\end{aligned}$$

et comme on a $\forall i = 1, ..., p,\ d_i < p$, alors :

$$-2\frac{\partial}{\partial b}\log(L) = 0 \quad \Leftrightarrow \quad b = \frac{\sum_{i=1}^{k} n_i \sum_{l=d_i+1}^{p}\lambda_{il}}{\sum_{i=1}^{k} n_i(p-d_i)},$$

ce qui permet de conclure. □

Estimation de α et σ : modèles $[\alpha\sigma Q_i d_i]$ et $[\alpha\sigma Q_i d]$ Les propositions 5.4 et 5.5 nous permettent de déduire les estimateurs de α et σ :

Corollaire 5.6. *Les estimateurs au sens du maximum de vraisemblance des paramètres α et σ existent et sont uniques :*

$$\hat{\alpha} = \frac{\hat{b}}{\hat{a}+\hat{b}},$$
$$\hat{\sigma^2} = \frac{\hat{a}\hat{b}}{\hat{a}+\hat{b}},$$

où $\hat{a}$ et $\hat{b}$ sont donnés aux propositions 5.4 et 5.5.

Estimation de α et σ_i : modèles $[\alpha\sigma_i Q_i d_i]$ et $[\alpha\sigma_i Q_i d]$

Proposition 5.7. *L'estimateur au sens du maximum de vraisemblance du paramètre α s'exprime en fonction des σ_i de la façon suivante :*

$$\hat{\alpha}(\sigma_1, ..., \sigma_k) = \frac{(\Lambda + np) - \sqrt{\Delta}}{2\Lambda},$$

avec les notations :

$$\begin{aligned}
\Delta &= (\Lambda + np)^2 - 4\Lambda\gamma, \\
\gamma &= \sum_{i=1}^{k} n_i d_i, \\
\Lambda &= \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \left(\sum_{l=1}^{d_i} \lambda_{il} - \sum_{l=d_i+1}^{p} \lambda_{il} \right),
\end{aligned}$$

et l'estimateur au sens du maximum de vraisemblance du paramètre σ_i^2 s'exprime en fonction de α de la façon suivante :

$$\forall i = 1, ..., k,\ \hat{\sigma_i^2}(\alpha) = \frac{1}{p} \left(\alpha \sum_{l=1}^{d_i} \lambda_{il} + (1-\alpha) \sum_{l=d_i+1}^{p} \lambda_{il} \right).$$

Démonstration. On a comme précédemment :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} \lambda_{il} \right),$$

avec $\delta_{il} = \frac{\sigma_i^2}{\alpha}$ si $l \leq d_i$ et $\frac{\sigma_i^2}{(1-\alpha)}$ sinon. Par conséquent, on peut écrire d'une part :

$$\begin{aligned}
-2\log(L) &= \sum_{i=1}^{k} n_i \left[\sum_{l=1}^{d_i} \left(2\log\sigma_i - \log\alpha + \frac{\alpha}{\sigma_i^2}\lambda_{il} \right) + \sum_{l=d_i+1}^{p} \left(2\log\sigma_i - \log(1-\alpha) + \frac{(1-\alpha)}{\sigma_i^2}\lambda_{il} \right) \right], \\
&= \sum_{i=1}^{k} n_i \left(2p\log\sigma_i - d_i\log\alpha - (p-d_i)\log(1-\alpha) + \frac{\alpha}{\sigma_i^2} \sum_{l=1}^{d_i} \lambda_{il} + \frac{(1-\alpha)}{\sigma_i^2} \sum_{l=d_i+1}^{p} \lambda_{il} \right).
\end{aligned}$$

Alors :

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\log(L) = 0 &\Leftrightarrow \sum_{i=1}^{k} n_i \left(-\frac{d_i}{\alpha} + \frac{(p-d_i)}{(1-\alpha)} + \frac{\sum_{l=1}^{d_i}\lambda_{il}}{\sigma_i^2} - \frac{\sum_{l=d_i+1}^{p}\lambda_{il}}{\sigma_i^2} \right) = 0, \\
&\Leftrightarrow -\frac{\gamma}{\alpha} + \frac{np-\gamma}{(1-\alpha)} + \Lambda = 0,
\end{aligned}$$

où $\gamma = \sum_{i=1}^{k} n_i d_i$ et $\Lambda = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \left(\sum_{l=1}^{d_i} \lambda_{il} - \sum_{l=d_i+1}^{p} \lambda_{il} \right)$. Donc,

$$\frac{\partial}{\partial \alpha} \log(L) = 0 \quad \Leftrightarrow \quad \psi(\alpha) = \Lambda \alpha^2 - (\Lambda + np)\alpha + \gamma = 0$$

On a :

$$\begin{aligned} \Delta &= (\Lambda + np)^2 - 4\Lambda\gamma, \\ &= \left(\Lambda + np(1 - 2\frac{\gamma}{np}) \right)^2 + (np)^2 \left(4\frac{\gamma}{np}(1 - \frac{\gamma}{np}) \right), \end{aligned}$$

or $\frac{\gamma}{np} < 1$ et par conséquent $\Delta > 0$. En remarquant que $\psi(0) = \gamma > 0$ et $\psi(1) = \gamma - np < 0$, on peut conclure qu'il existe une unique solution dans $[0, 1]$: la plus petite des deux. D'autre part,

$$\begin{aligned} \frac{\partial}{\partial \sigma_i} \log(L) = 0 \quad &\Leftrightarrow \quad n_i \left(\frac{2p}{\sigma_i} - \frac{2\alpha \sum_{l=1}^{d_i} \lambda_{il}}{\sigma_i^3} - \frac{2(1-\alpha) \sum_{l=d_i+1}^{p} \lambda_{il}}{\sigma_i^3} \right) = 0, \\ &\Leftrightarrow \quad p - \frac{1}{\sigma_i^2} \left(\alpha \sum_{l=1}^{d_i} \lambda_{il} + (1-\alpha) \sum_{l=d_i+1}^{p} \lambda_{il} \right) = 0, \end{aligned}$$

donc :

$$\sigma_i^2 = \frac{1}{p} \left(\alpha \sum_{l=1}^{d_i} \lambda_{il} + (1-\alpha) \sum_{l=d_i+1}^{p} \lambda_{il} \right).$$

□

Estimation de α_i et σ : modèles $[\alpha_i \sigma Q_i d_i]$ et $[\alpha_i \sigma Q_i d]$

Proposition 5.8. *L'estimateur au sens du maximum de vraisemblance du paramètre σ s'exprime en fonction des α_i de la façon suivante :*

$$\hat{\sigma^2}(\alpha_1, ..., \alpha_k) = \frac{1}{np} \sum_{i=1}^{k} n_i \left(\alpha_i \sum_{l=1}^{d_i} \lambda_{il} + (1 - \alpha_i) \sum_{l=d_i+1}^{p} \lambda_{il} \right),$$

et l'estimateur au sens du maximum de vraisemblance du paramètre α_i s'exprime en fonction de σ de la façon suivante :

$$\forall i = 1, ..., k, \; \hat{\alpha_i}(\sigma^2) = \frac{(\Lambda_i + p) - \sqrt{\Delta_i}}{2\Lambda_i},$$

avec les notations :

$$\begin{aligned} \Delta_i &= (\Lambda_i + p)^2 - 4\Lambda_i d_i, \\ \Lambda_i &= \frac{\sum_{l=1}^{d_i} \lambda_{il} - \sum_{l=d_i+1}^{p} \lambda_{il}}{\sigma^2}. \end{aligned}$$

Démonstration. On a comme précédemment :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}} \lambda_{il} \right),$$

avec $\delta_{il} = \frac{\sigma^2}{\alpha_i}$ si $l \leq d_i$ et $\frac{\sigma^2}{(1-\alpha_i)}$ sinon. Par conséquent, on peut écrire :

$$-2\log(L) = \sum_{i=1}^{k} n_i \left[\sum_{l=1}^{d_i} \left(2\log\sigma - \log\alpha_i + \frac{\alpha_i}{\sigma^2} \lambda_{il} \right) + \sum_{l=d_i+1}^{p} \left(2\log\sigma - \log(1-\alpha_i) + \frac{(1-\alpha_i)}{\sigma^2} \lambda_{il} \right) \right],$$

$$= \sum_{i=1}^{k} n_i \left(2p \log \sigma - d_i \log \alpha_i - (p-d_i) \log(1-\alpha_i) + \frac{\alpha_i}{\sigma^2} \sum_{l=1}^{d_i} \lambda_{il} + \frac{(1-\alpha_i)}{\sigma^2} \sum_{l=d_i+1}^{p} \lambda_{il} \right).$$

Alors :

$$\frac{\partial}{\partial \sigma} \log(L) = 0 \quad \Leftrightarrow \quad \frac{2np}{\sigma} - \frac{2}{\sigma^3} \sum_{i=1}^{k} n_i \left(\alpha_i \sum_{l=1}^{d_i} \lambda_{il} + (1-\alpha_i) \sum_{l=d_i+1}^{p} \lambda_{il} \right) = 0,$$

ce qui permet d'obtenir l'expression de l'estimateur de σ^2 en fonction des α_i. D'autre part,

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} \log(L) = 0 \quad &\Leftrightarrow \quad n_i \left(-\frac{d_i}{\alpha_i} + \frac{(p-d_i)}{(1-\alpha_i)} + \frac{1}{\sigma^2} \sum_{l=1}^{d_i} \lambda_{il} - \frac{1}{\sigma^2} \sum_{l=d_i+1}^{p} \lambda_{il} \right) = 0, \\ &\Leftrightarrow \quad p\alpha_i - d_i + \frac{\alpha_i(1-\alpha_i)}{\sigma^2} \left(\sum_{l=1}^{d_i} \lambda_{il} - \sum_{l=d_i+1}^{p} \lambda_{il} \right) = 0, \\ &\Leftrightarrow \quad \psi_i(\alpha_i) = \alpha_i^2 \Lambda_i - (\Lambda_i + p)\alpha_i + d_i = 0, \end{aligned}$$

avec $\Lambda_i = \frac{\sum_{l=1}^{d_i} \lambda_{il} - \sum_{l=d_i+1}^{p} \lambda_{il}}{\sigma^2}$. On a :

$$\begin{aligned} \Delta_i &= (\Lambda_i + p)^2 - 4\Lambda_i d_i, \\ &= (\Lambda_i + p(1 - 2\frac{d_i}{p}))^2 + p^2 \left(4\frac{d_i}{p}(1 - \frac{d_i}{p}) \right), \end{aligned}$$

or $\frac{d_i}{p} < 1$ et par conséquent $\Delta_i > 0$.En remarquant que $\psi_i(0) = d_i > 0$ et $\psi_i(1) = d_i - p < 0$, on peut conclure qu'il existe une unique solution $\in [0,1]$: la plus petite des deux. □

Remarque 4. Les estimateurs des propositions 5.7 et 5.8 n'ayant pas de formulation explicite, ils doivent être calculés grâce à une procédure itérative. On pourra choisir d'initialiser la procédure avec les valeurs de la proposition 5.3 et utiliser par exemple une procédure telle que celle présentée ci-dessous pour le calcul des estimateurs du modèle $[\alpha \sigma_i Q_i d_i]$:

– Initialisation :
$\forall i = 1, ..., k$, $\sigma_i^2(0) = \hat{\sigma_i^2}$,
$\alpha(0) = \hat{\alpha}(\sigma_1(0), ..., \sigma_k(0))$,
$j = 0$.

– Actualisation : jusqu'à convergence, faire
$\forall i = 1, ..., k$, $\sigma_i^2(j+1) = \frac{1}{p} \left(\alpha(j) \sum_{l=1}^{d_i} \lambda_{il} + (1 - \alpha(j)) \sum_{l=d_i+1}^{p} \lambda_{il} \right)$,
$\alpha(j+1) = \hat{\alpha}(\sigma_1(j+1), ..., \sigma_k(j+1))$.
$j = j + 1$,

5.4 Estimateurs des règles particulières à Q_i communs

Nous présentons dans ce paragraphe les estimateurs des règles particulières ayant un modèle à Q_i communs, *i.e.* $\forall i = 1, ..., k$, $Q_i = Q$.

Estimation de a et b : modèle $[abQd]$

Proposition 5.9. *Les estimateurs au sens du maximum de vraisemblance de la matrice Q et des paramètres a et b existent et sont uniques :*

(i) Q est estimée par la matrice $\hat{Q}$ dont les d premières colonnes sont les vecteurs propres associés aux d plus grandes valeurs propres de $\hat{W} = \sum_{i=1}^{k} \pi_i \hat{\Sigma}_i$ et les $(p-d)$ dernières colonnes sont les vecteurs propres associés aux $(p-d)$ plus petites valeurs propres de $\hat{W}$,

(ii) a *est estimé par la moyenne des* d *plus grandes valeurs propres de* $\hat{W}$ *:*

$$\hat{a} = \frac{\sum_{l=1}^{d} \lambda_l}{d},$$

(iii) et b *est estimé par la moyenne des* $(p-d)$ *plus petites valeurs propres de* $\hat{W}$ *:*

$$\hat{b} = \frac{\sum_{l=d+1}^{p} \lambda_l}{(p-d)},$$

où λ_l *est la* $l^{\text{ème}}$ *plus grande valeur propre de* $\hat{W}$*.*

Démonstration. Le lemme 5.1 nous permet d'écrire :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p} \left(\log \delta_l + \frac{1}{\delta_l} q_l^t \hat{\Sigma}_i q_l \right) + C^{te},$$

avec $\delta_l = a$ si $l \leq d$ et $\delta_l = b$ sinon. On souhaite minimiser cette quantité sous la contrainte $q_l^t q_l = 1$, ce qui revient à minimiser la fonction de Lagrange suivante :

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{k} n_i \sum_{l=1}^{p} \left(\log \delta_l + \frac{1}{\delta_l} q_l^t \hat{\Sigma}_i q_l \right) - \sum_{l=1}^{p} \theta_l (q_l^t q_l - 1), \\ &= n \sum_{l=1}^{p} \log \delta_l + \sum_{l=1}^{p} \frac{1}{\delta_l} q_l^t \left(\sum_{i=1}^{k} n_i \hat{\Sigma}_i \right) q_l - \sum_{l=1}^{p} \theta_l (q_l^t q_l - 1), \end{aligned}$$

où les θ_l sont les multiplicateurs de Lagrange. Or, $\sum_{i=1}^{k} n_i \hat{\Sigma}_i$ n'est autre que n fois l'estimateur de la matrice de variance intra-classe W. Par conséquent, on souhaite minimiser la fonction suivante :

$$\mathcal{L} = n \sum_{l=1}^{p} \left(\log \delta_l + \frac{1}{\delta_l} q_l^t \hat{W} q_l \right) - \sum_{l=1}^{p} \theta_l (q_l^t q_l - 1).$$

La dérivée partielle de $\mathcal{L}$ par rapport à a vaut :

$$\frac{\partial \mathcal{L}}{\partial a} = n \sum_{l=1}^{p} \frac{\partial}{\partial \delta_l} \left(\log \delta_l + \frac{1}{\delta_l} q_l^t \hat{W} q_l \right) \frac{\partial \delta_l}{\partial a},$$

avec $\frac{\partial \delta_l}{\partial a} = 1$ si $l \leq d$ et 0 sinon. Par conséquent,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial a} &= n \sum_{l=1}^{d} \left(\frac{1}{a} - \frac{1}{a^2} q_l^t \hat{W} q_l \right), \\ &= \frac{nd}{a} - \frac{n}{a^2} \sum_{l=1}^{d} q_l^t \hat{W} q_l, \end{aligned}$$

et la condition $\frac{\partial \mathcal{L}}{\partial a} = 0$ implique que :

$$\hat{a} = \frac{1}{d} \sum_{l=1}^{d} q_l^t \hat{W} q_l. \tag{5.5}$$

De même, la dérivée partielle de $\mathcal{L}$ par rapport à b vaut :

$$\frac{\partial \mathcal{L}}{\partial b} = n \sum_{l=1}^{p} \frac{\partial}{\partial \delta_l} \left(\log \delta_l + \frac{1}{\delta_l} q_l^t \hat{W} q_l \right) \frac{\partial \delta_l}{\partial b},$$

avec $\frac{\partial \delta_l}{\partial b} = 1$ si $l \geq d+1$ et 0 sinon. Donc

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b} &= n \sum_{l=d+1}^{p} \left(\frac{1}{b} - \frac{1}{b_l^2} q_l^t \hat{W} q_l \right), \\ &= \frac{n(p-d)}{b} - \frac{n}{b^2} \sum_{l=d+1}^{p} q_l^t \hat{W} q_l,\end{aligned}$$

et la condition $\frac{\partial \mathcal{L}}{\partial b} = 0$ implique que :

$$\hat{b} = \frac{1}{(p-d)} \sum_{l=d+1}^{p} q_l^t \hat{W} q_l. \tag{5.6}$$

Enfin, le gradient de $\mathcal{L}$ par rapport à q_l vaut :

$$\nabla_{q_l} \mathcal{L} = 2\frac{n}{\delta_l} \hat{W} q_l - 2\theta_l q_l,$$

et en multipliant cette quantité à gauche par q_l^t, on a :

$$\begin{aligned} q_l^t \nabla_{q_l} \mathcal{L} = 0 \quad &\Leftrightarrow \quad 2\frac{n}{\delta_l} q_l^t \hat{W} q_l - 2\theta_l = 0, \\ &\Leftrightarrow \quad \theta_l = \frac{n}{\delta_l} q_l^t \hat{W} q_l, \end{aligned}$$

et par conséquent :

$$\hat{W} q_l = \frac{\theta_l \delta_l}{n} q_l,$$

ce qui signifie que q_l est le vecteur propre de $\hat{W}$ associé à la valeur propre $\lambda_l = \frac{\theta_l \delta_l}{n}$. En reportant dans les expressions (5.5) et (5.6), cela permet d'établir les résultats (i) et (ii). Les q_l étant vecteurs propres de $\hat{W}$ qui est une matrice symétrique, cela implique que $q_l^t q_h = 0$ si $h \neq l$ et que $q_l^t q_l = 1$. Il ne nous reste plus qu'à trouver l'ordre des vecteurs propres de $\hat{W}$ dans Q. On souhaite minimiser la quantité suivante à l'*optimum* :

$$-2 \log L = n(d \log \hat{a} + (p-d) \log \hat{b}),$$

avec $\hat{a} > \hat{b}$ et $d\hat{a} + (p-d)\hat{b} = \sum_{l=1}^{p} \lambda_l = s$ qui est la trace de $\hat{W}$ et donc $\hat{a} > \frac{s}{p}$. Alors,

$$\frac{-2\partial \log L}{\partial \hat{a}} = \frac{d}{\hat{a}} + \frac{p-d}{s - d\hat{a}} < 0,$$

et donc, il faut choisir $\hat{a}$ le plus grand possible pour minimiser $-2 \log L$. Cela ne peut être réalisé qu'en choisissant les vecteurs propres associés aux d plus grandes valeurs propres de $\hat{W}$ pour remplir les d premières colonnes de Q et, par conséquent, en choisissant les vecteurs propres associés aux $(p-d)$ plus petites valeurs propres de $\hat{W}$ pour remplir les $(p-d)$ dernières colonnes de Q □

Estimation de α et σ_i : modèle $[\alpha \sigma_i Q d]$

Proposition 5.10. *Les estimateurs au sens du maximum de vraisemblance de la matrice Q et des paramètres α et σ_i existent :*

(i) l'estimateur au sens du maximum de vraisemblance de Q est la matrice $\hat{Q}(\sigma_1, ..., \sigma_k)$ dont les d premières colonnes sont les vecteurs propres associés aux d plus grandes valeurs propres de $S(\sigma_1, ..., \sigma_k)$ définie par :

$$S(\sigma_1, ..., \sigma_k) = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \hat{\Sigma}_i$$

et les $(p-d)$ dernières colonnes sont les vecteurs propres associés aux $(p-d)$ plus petites valeurs propres de $S(\sigma_1, ..., \sigma_k)$,

(ii) l'estimateur au sens du maximum de vraisemblance de α s'exprime en fonction de Q et des σ_i^2 de la façon suivante :

$$\hat{\alpha}(\sigma_1,...,\sigma_k,Q) = \frac{(\Lambda+np)-\sqrt{(\Lambda+np)^2-4nd\Lambda}}{2\Lambda},$$

où $\Lambda(\sigma_1,...,\sigma_k) = \sum_{l=1}^{d} q_l^t S(\sigma_1,...,\sigma_k) q_l - \sum_{l=d+1}^{p} q_l^t S(\sigma_1,...,\sigma_k) q_l$,

(iii) et l'estimateur au sens du maximum de vraisemblance de σ_i^2 s'exprime en fonction de α et de Q de la façon suivante :

$$\forall i = 1,...,k,\ \hat{\sigma_i^2}(\alpha,Q) = \frac{1}{p}\left(\alpha\sum_{l=1}^{d} q_l^t\hat{\Sigma}_i q_l + (1-\alpha)\sum_{l=d+1}^{p} q_l^t\hat{\Sigma}_i q_l\right).$$

Démonstration. Le lemme 5.1 nous permet d'écrire :

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p}\left(\log\delta_{il} + \frac{1}{\delta_{il}} q_l^t\hat{\Sigma}_i q_l\right) + C^{te},$$

avec $\delta_{il} = \frac{\sigma_i^2}{\alpha}$ si $l \leq d$ et $\delta_{il} = \frac{\sigma_i^2}{1-\alpha}$ sinon. On obtient donc :

$$\begin{aligned}
-2\log(L) &= \sum_{i=1}^{k} n_i \sum_{l=1}^{d}\left(\log(\frac{\sigma_i^2}{\alpha}) + \frac{\alpha}{\sigma_i^2} q_l^t\hat{\Sigma}_i q_l\right) + \sum_{i=1}^{k} n_i \sum_{l=d+1}^{p}\left(\log(\frac{\sigma_i^2}{1-\alpha}) + \frac{1-\alpha}{\sigma_i^2} q_l^t\hat{\Sigma}_i q_l\right) + C^{te},\\
&= \alpha\sum_{i=1}^{k}\sum_{l=1}^{d}\frac{n_i}{\sigma_i^2} q_l^t\hat{\Sigma}_i q_l + (1-\alpha)\sum_{i=1}^{k}\sum_{l=d+1}^{p}\frac{n_i}{\sigma_i^2} q_l^t\hat{\Sigma}_i q_l\\
&\quad + d\left(\sum_{i=1}^{k} n_i\log(\sigma_i^2) - n\log\alpha\right) + (p-d)\left(\sum_{i=1}^{k} n_i\log(\sigma_i^2) - n\log(1-\alpha)\right) + C^{te},\\
&= \alpha\sum_{l=1}^{d} q_l^t\left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\hat{\Sigma}_i\right) q_l + (1-\alpha)\sum_{l=d+1}^{p} q_l^t\left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\hat{\Sigma}_i\right) q_l\\
&\quad + p\sum_{i=1}^{k} n_i\log(\sigma_i^2) - n\left(d\log\alpha + (p-d)\log(1-\alpha)\right) + C^{te}.
\end{aligned}$$

Soit $S(\sigma_1,...,\sigma_k) = \sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\hat{\Sigma}_i$, alors :

$$-2\log(L) = \alpha\sum_{l=1}^{d} q_l^t S q_l + (1-\alpha)\sum_{l=d+1}^{p} q_l^t S q_l + p\sum_{i=1}^{k} n_i\log(\sigma_i^2) - n\left(d\log\alpha + (p-d)\log(1-\alpha)\right) + C^{te}.$$

On souhaite minimiser cette quantité sous la contrainte $q_l^t q_l = 1$, ce qui revient à minimiser la fonction de Lagrange suivante :

$$\mathcal{L} = -2\log(L) - \sum_{l=1}^{p}\theta_l(q_l^t q_l - 1),$$

où les θ_{il} sont les multiplicateurs de Lagrange. Le gradient de $\mathcal{L}$ par rapport à q_l, $\forall l \leq d$, vaut :

$$\nabla_{q_l}\mathcal{L} = 2\left(\alpha S q_l - \theta_l q_l\right), \tag{5.7}$$

et en multipliant cette quantité à gauche par q_l^t, on a :

$$\begin{aligned}
q_l^t\nabla_{q_l}\mathcal{L} = 0 &\Leftrightarrow 2\alpha q_l^t S q_l - 2\theta_l = 0,\\
&\Leftrightarrow \theta_l = \alpha q_l^t S q_l,
\end{aligned}$$

et par conséquent :

$$Sq_l = \frac{\theta_l}{\alpha} q_l,$$

ce qui signifie que $\forall l \leq d$, q_l est le vecteur propre de S associé à la valeur propre $\lambda_l = \frac{\theta_l}{\alpha}$. En effectuant le même raisonnement pour $l > d$, on montre que $\forall l > d$, q_l est le vecteur propre de S associé à la valeur propre $\lambda_l = \frac{\theta_l}{1-\alpha}$. Les q_l étant vecteurs propres de S qui est une matrice symétrique, cela implique que $q_l^t q_h = 0$ si $h \neq l$ et que $q_l^t q_l = 1$. De manière similaire à la démonstration de la proposition 5.2, on montre que les d premières colonnes de Q sont les vecteurs propres associés aux d plus grandes valeurs propres de S et les $(p-d)$ dernières colonnes sont les vecteurs propres associés aux $(p-d)$ plus petites valeurs propres car $\frac{\sigma_i^2}{\alpha} > \frac{\sigma_i^2}{1-\alpha}$ par hypothèse. D'autre part, la dérivée partielle de $\mathcal{L}$ par rapport à σ_i vaut :

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = \frac{2pn_i}{\sigma_i} - \alpha \sum_{l=1}^{d} q_l^t \frac{2n_i}{\sigma_i^3} \hat{\Sigma}_i q_l - (1-\alpha) \sum_{l=d+1}^{p} q_l^t \frac{2n_i}{\sigma_i^3} \hat{\Sigma}_i q_l.$$

Alors, on a :

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = 0 \quad \Leftrightarrow \quad \sigma_i^2 = \frac{1}{p} \left(\alpha \sum_{l=1}^{d} q_l^t \hat{\Sigma}_i q_l + (1-\alpha) \sum_{l=d+1}^{p} q_l^t \hat{\Sigma}_i q_l \right).$$

Ce qui nous permet d'exprimer σ_i^2 en fonction de α et des q_l. Enfin, la dérivée partielle de $\mathcal{L}$ par rapport à α vaut :

$$\frac{\partial \mathcal{L}}{\partial \alpha} = -\frac{nd}{\alpha} + \frac{n(p-d)}{1-\alpha} + \sum_{l=1}^{d} q_l^t S q_l - \sum_{l=d+1}^{p} q_l^t S q_l.$$

Alors, on a :

$$\frac{\partial \mathcal{L}}{\partial \alpha} = 0 \quad \Leftrightarrow \quad \psi(\alpha) = \Lambda \alpha^2 - (\Lambda + np)\alpha + nd = 0,$$

avec $\Lambda = \sum_{l=1}^{d} q_l^t S q_l - \sum_{l=d+1}^{p} q_l^t S q_l$. On a :

$$\begin{aligned} \Delta &= (\Lambda + np)^2 - 4\Lambda nd, \\ &= (\Lambda + np(1 - 2\frac{d}{p}))^2 + (np)^2 \left(4\frac{d}{p}(1 - \frac{d}{p}) \right), \end{aligned}$$

or $\frac{d}{p} < 1$ et par conséquent $\Delta > 0$. En remarquant que $\psi(0) = nd > 0$ et $\psi(1) = n(d-p) < 0$, on peut conclure qu'il existe une unique solution $\in [0, 1]$: la plus petite des deux. □

Remarque 5. Les estimateurs de la proposition 5.10 n'ayant pas de formulation explicite, ils doivent être calculés grâce à une procédure itérative. On pourra choisir d'initialiser la procédure avec les valeurs de la proposition 5.3. On pourra utiliser par exemple une procédure telle que celle présentée ci-dessous :

– Initialisation :
  $\forall i = 1, ..., k$, $\sigma_i^2(0) = \hat{\sigma_i^2}$,
  $\alpha(0) = \hat{\alpha}(\sigma_1(0), ..., \sigma_k(0))$ et $Q(0) = \hat{Q}(\sigma_1(0), ..., \sigma_k(0))$.
  $j = 0$.
– Actualisation : jusqu'à convergence, faire
  $\forall i = 1, ..., k$, $\sigma_i^2(j+1) = \frac{1}{p} \left(\alpha(j) \sum_{l=1}^{d} q_l^t(j) \hat{\Sigma}_i q_l(j) + (1 - \alpha(j)) \sum_{l=d+1}^{p} q_l^t(j) \hat{\Sigma}_i q_l(j) \right)$,
  $\alpha(j+1) = \hat{\alpha}(\sigma_1(j+1), ..., \sigma_k(j+1))$,
  $j = j + 1$.

Autres modèles à Q_i communs Dans ces autres cas, les estimateurs du maximum de vraisemblance des différents paramètres n'ont pas de forme explicite et doivent donc être déterminés grâce une procédure itérative. Cette procédure itérative est basée sur l'algorithme FG [11].

5.5 Estimation de la dimension intrinsèque

La démarche que nous avons adoptée dans ce rapport fait l'hypothèse que les données de la classe C_i, $\forall i = 1, ..., k$, vivent dans un espace de dimension intrinsèque d_i. Par conséquent, il nous faut à présent déterminer le paramètre d_i.
La détermination de la dimension intrinsèque d'un jeu de données est un problème difficile qui ne possède pas de solution explicite. De nombreuses méthodes ont été proposées pour estimer cette dimension intrinsèque, mais aucune ne permet de résoudre efficacement le problème. Nous nous proposons d'utiliser deux méthodes empiriques pour trouver la valeur de d_i à partir de la base d'apprentissage A.

Estimation de d_i par seuillage commun sur la variance cumulée L'idée naturelle pour estimer la dimension de l'espace propre $\mathbb{E}_i$ est d'utiliser les valeurs propres de la matrice de variance Σ_i, étant donné qu'elles sont caractéristiques de la variance des données projetées. En effet, la i^e valeur propre de Σ_i correspond au pourcentage de variance porté par le i^e vecteur propre. Par conséquent, on peut rechercher, par seuillage sur la variance cumulée de la classe C_i, la dimension $\hat{d}_i$:

$$\hat{d}_i = \underset{d=1,\ldots,p-1}{\operatorname{argmin}} \left\{ \frac{\sum_{j=1}^{d} \lambda_{ij}}{\sum_{j=1}^{p} \lambda_{ij}} \geq s \right\},$$

où $s \in [0, 1]$ est le seuil commun et λ_{ij} est la j^e valeur propre de Σ_{ij}. Cette stratégie revient à faire une ACP par classes. Le s choisi est le seuil maximisant le taux de classification correcte des données d'apprentissage (voir chapitre 6).

Remarque 6. Le fait de choisir un seuil commun sur la variance cumulée de chaque classe n'implique généralement pas que les d_i sont égaux.

Estimation du d_i indépendemment pour chaque classe On peut également vouloir trouver la dimension d_i « optimale » pour chaque classe C_i indépendemment. Avec les hypothèses de notre modèle, le $\hat{d}_i^*$ est la dimension qui va donner une fonction de coût minimale pour les éléments de la classe C_i et une fonction de coût maximale pour les éléments des autres classes, *i.e.* :

$$\hat{d}_i^* = \underset{d=1,\ldots,p-1}{\operatorname{argmin}} \{ \frac{\sum_{x_j \in C_i} K_i(x_j)}{\sum_{x_l \notin C_i} K_i(x_l)} \}, \tag{5.8}$$

où $K_i(x) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i) + C^{te}$.

Estimation de $d_i = d$ Toutefois, si l'on a de bonnes raisons de penser que la dimension des espaces dans lesquelles vivent les données sont égales, alors il n'y a qu'une seule dimension d à estimer. On peut alors choisir d'estimer d par la dimension qui maximise le taux de classification correcte de l'ensemble des données d'apprentissage.

6 Résultats expérimentaux

Afin de vérifier le bien-fondé de la méthode que nous proposons dans ce rapport, nous l'avons mise en œuvre et comparée sur des données synthétiques et réelles. Les comparaisons rapportées dans ces lignes ont été faites avec des méthodes classiques que nous avons estimées être de référence.

6.1 Algorithme et protocole

Nous allons présenter dans ces lignes le protocole de mise en œuvre et l'algorithme de l'HDDA que nous avons utilisé. L'utilisation de nos méthodes se déroule en deux phases :

– Apprentissage du seuil de dimensionnalité :
  Pour tous les seuils $s \in]0, 1[$, on estime le taux de classification correcte par validation croisée. C'est à dire que l'on applique n fois l'algorithme du tableau 6.1 avec des échantillons d'apprentissage de taille $n - 1$. Le seuil $\hat{s}$ retenu est celui qui maximisent le taux de classification correcte.

Entrées : Données d'apprentissage, données à classer, seuil de dimensionnalité,

Sorties : classes des données, probabilité d'erreur de classement.

(i) **Calcul des estimateurs et de** K_i : Pour chaque classe : $i = 1, ..., k$

(a) Calcul des valeurs et vecteurs propres de la matrice de variance Σ_i,

(b) Détermination de la dimension intrinsèque de l'espace propre $\mathbb{E}_i$,

(c) Calcul des estimateurs $\hat{a_i}$ et $\hat{b_i}$,

(d) Projection du point x à classer dans l'espace propre $\mathbb{E}_i$,

(e) Calcul de la fonction de coût $K_i(x)$.

(ii) **Classification** : $x \in C_{i^*}$ si $i^* = \text{argmin}_{i=1,...,k}\{K_i(x)\}$.

TAB. 6.1 – Algorithme de l'Analyse Discriminante de Haute Dimension (modèle $[a_i b_i Q_i d_i]$).

– Classification :
Pour le seuil $\hat{s}$ obtenu, on peut alors classer les données grâce à l'implantation de l'Analyse Discriminante de Haute Dimension du tableau 6.1.

Nous avons choisi de comparer les méthodes de discrimination suivantes :

(i) Méthodes d'Analyse Discriminante de Haute Dimension :
– 14 des 24 modèles présentés au tableau 4.1.

(ii) Méthodes d'Analyse Discriminante :
– Analyse Discriminante Quadratique (QDA),
– Analyse Discriminante Linéaire (LDA),
– Analyse Factorielle Discriminante (FDA).

(iii) Méthodes à noyaux :
– *Support Vector Machine* (SVM) à noyau gaussien [14, Chap. 12].

6.2 Les données

Avant de mettre en œuvre notre méthode sur des données réelles dont on ne connaît pas nécessairement la nature statistique, nous l'avons testé sur des données synthétiques et sur un jeu de données de référence : les données Iris de Fisher. Nous l'avons ensuite utiliser pour classifier un jeu de données issue du domaine de la vison par ordinateur. Ce jeu de données (LIS) correspond à une application de catégorisation d'images naturelles. Nous donnons ci-dessous la nature et l'origine des différents jeux de données.

Données synthétiques Nous avons simulé trois densités gaussiennes différentes vivant respectivement dans des espaces de dimension $d_1 = 3$, $d_2 = 4$ et $d_3 = 5$ et plongées dans $\mathbb{R}^{15}$. Il est a noter que les dimensions dans lesquelles vivent les éléments des trois classes se chevauchent ce qui rend plus difficile la tâche de classification. Le jeu de données ainsi créé comporte 500 vecteurs en dimension 15 répartis en 3 classes. Nous avons choisi de nous placer dans un cas où les proportions des classes sont différentes : $\pi_1 = \frac{1}{2}$, $\pi_2 = \frac{1}{3}$ et $\pi_3 = \frac{1}{6}$ (voir Fig. 6.1).

Données Iris de Fisher Les données Iris de Fisher, initialement publiées par Fischer [9], est un jeu de données de référence dans le domaine de la classification. Le problème de la classification de ces données et particulièrement intéressant car une classe est linéairement séparable des deux autres, mais les deux dernières ne le sont pas. Nous avons choisit d'appliquer notre méthode à cet exemple car il est fréquemment utilisé et facilement disponible. Le jeu de données comporte 150 exemples en dimension 4 équirépartis en 3 classes : Iris *setosa*, *versicolor* et *virginica*.

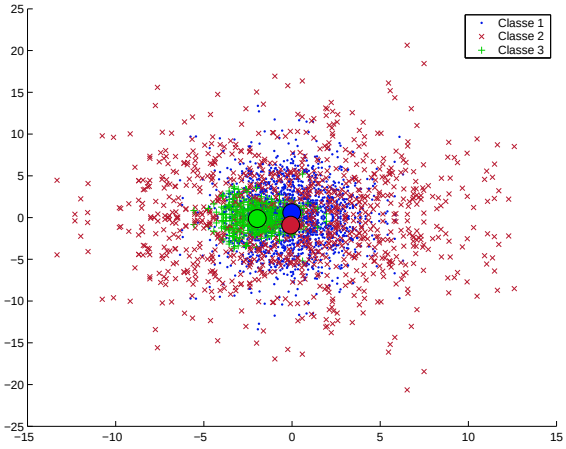


FIG. 6.1 – Données synthétiques : projection des trois densités gaussiennes simulées sur les 2 axes discriminants de l'AFD.

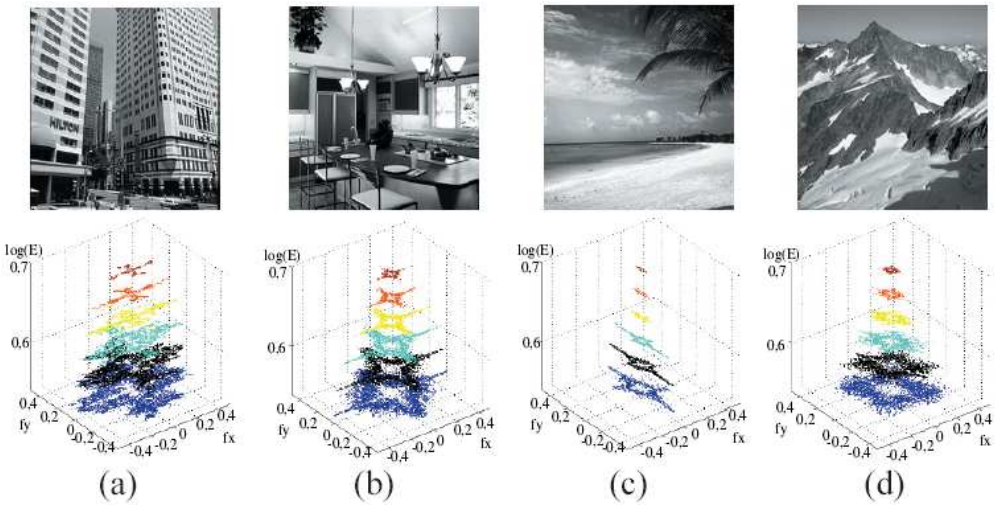


FIG. 6.2 – Données LIS : Réponses énergétiques aux filtres de Gabor d'images (a) de villes, (b) d'intérieurs, (c) de plages et (d) de montagnes (figure issue de [16]).

Taux de classification correcte	$[Qd]$	$[Qd_i]$	$[Q_id]$	$[Q_id_i]$
Modèle $[ab]$	0.538 ($d = 3$)	/	0.7 ($d = 3$)	0.746 ($s = 0.75$)
Modèle $[a_ib]$	/	/	0.858 ($d = 14$)	0.874 ($s = 0.75$)
Modèle $[ab_i]$	/	/	0.934 ($d = 3$)	0.866 ($s = 0.75$)
Modèle $[\alpha_i\sigma]$	/	/	0.82 ($d = 14$)	0.82 ($s = 0.99$)
Modèle $[\alpha\sigma_i]$	0.626 ($d = 12$)	/	0.832 ($d = 3$)	0.802 ($s = 0.77$)
Modèle $[a_ib_i]$	/	/	**0.964** ($d = 3$)	0.958 ($s = 0.82$)
Méthodes de référence	QDA 0.942	LDA 0.512	FDA 0.51	SVM 0.478

TAB. 6.2 – Résultats de classification pour les données synthétiques.

Données LIS Ces données ont été obtenues en se basant sur un modèle d'inspiration biologique de description des images [16, 17]. A chaque image correspond un vecteur (descripteur) de dimension 49 ; chacune des dimensions est la valeur de la réponse énergétique de l'image à un filtre de Gabor pour certaines fréquences et orientations. La figure 6.2 présente les réponses énergétiques aux filtres de Gabor d'images de différentes natures. Le but est donc de catégoriser des images représentées par des descripteurs en grande dimension. On peut montrer simplement, en utilisant l'ACP, que la réduction de dimension de ces données augmente le taux de bonne catégorisation. On peut donc raisonnablement penser que notre méthode va également permettre d'augmenter ce taux. Nous présentons dans la suite les résultats obtenus avec notre méthode et nous les comparons à des méthodes classiques. Le jeu de données LIS comporte 328 descripteurs en dimension 49 répartis en 4 classes : images de plages, de villes, de montagnes et d'intérieurs. Les proportions de chacune des classes sont égales et valent $\pi_i = \frac{1}{4}$, $\forall i = 1, ..., 4$.

6.3 Résultats et discussion

Pour la méthode SVM, nous avons utilisé un noyau gaussien avec les paramètres par défaut, mais nous avons remarqué que le fait de changer de noyau ou de modifier les paramètres n'avait que peu d'incidences sur les résultats. Nous présentons toutefois les meilleurs résultats que nous avons obtenus avec cette méthode. Nous commenterons ensuite les résultats numériques et nous montrerons quels sont les principaux avantages de notre méthode.

Données synthétiques Le tableau 6.2 présente les résultats de classification obtenus sur le jeu de données synthétiques. La mise en œuvre de l'Analyse Discriminante de Haute Dimension sur les données synthétiques nous a permis de vérifier que notre méthode est tout à fait adapté aux données de grande dimension vivant dans des espaces de dimension intrinsèque inférieure. La comparaison des résultats de l'HDDA avec ceux obtenus avec des méthodes de références permet d'apprécier la difficulté de la classification de ce jeu de données artificiel. Il est naturel que la méthode QDA donne des résultats satisfaisants car la nature des données est aussi adaptée à cette méthode. L'HDDA s'est également révélée particulièrement robuste à la différence de proportion entre les classes. D'autre part, cette expérience a mis en évidence la rapidité de calcul de l'HDDA : le temps nécessaire à classifier les 500 individus est de l'ordre de 0.04 secondes pour

Taux de classification correcte	$[Qd]$	$[Qd_i]$	$[Q_id]$	$[Q_id_i]$
Modèle $[ab]$	0.987 ($d=1$)	/	0.98 ($d=1$)	0.98 ($s=0.75$)
Modèle $[a_ib]$	/	/	0.973 ($d=1$)	**0.993** ($s=0.9$)
Modèle $[ab_i]$	/	/	0.98 ($d=1$)	0.987 ($s=0.89$)
Modèle $[\alpha_i\sigma]$	/	/	0.973 ($d=1$)	**0.993** ($s=0.9$)
Modèle $[\alpha\sigma_i]$	0.96 ($d=3$)	/	0.973 ($d=1$)	0.973 ($s=0.75$)
Modèle $[a_ib_i]$	/	/	0.973 ($d=1$)	**0.993** ($s=0.9$)
Méthodes de référence	QDA 0.973	LDA 0.98	FDA 0.98	SVM 0.967

TAB. 6.3 – Résultats de classification pour les données Iris de Fisher.

Taux de classification correcte	$[Qd]$	$[Qd_i]$	$[Q_id]$	$[Q_id_i]$
Modèle $[ab]$	0.78 ($d=4$)	/	0.878 ($d=28$)	0.848 ($s=0.97$)
Modèle $[a_ib]$	/	/	**0.881** ($d=29$)	0.851 ($s=0.98$)
Modèle $[ab_i]$	/	/	0.872 ($d=28$)	0.86 ($s=0.98$)
Modèle $[\alpha_i\sigma]$	/	/	**0.881** ($d=29$)	0.851 ($s=0.97$)
Modèle $[\alpha\sigma_i]$	0.756 ($d=27$)	/	0.875 ($d=29$)	0.845 ($s=0.97$)
Modèle $[a_ib_i]$	/	/	0.872 ($d=28$)	0.857 ($s=0.98$)
Méthodes de référence	QDA 0.849	LDA 0.775	FDA 0.79	SVM 0.839

TAB. 6.4 – Résultats de classification pour les données LIS.

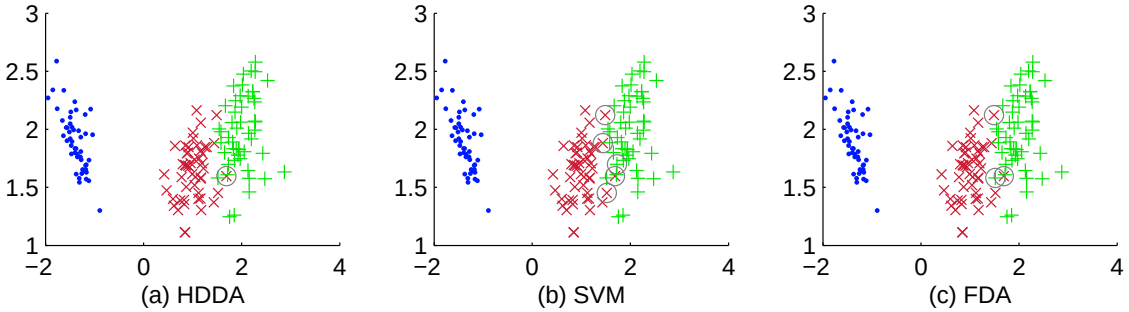


FIG. 6.3 – Projection sur les deux premiers axes discriminants du résultat de la classification des données Iris de Fisher avec les classifieurs (a) HDDA, (b) SVM et (c) FDA. Les erreurs de classification sont encerclées.

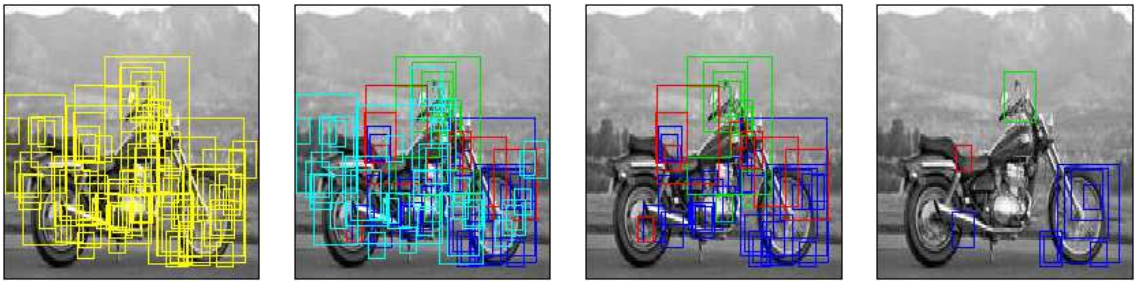
l'HDDA et les méthodes d'Analyse Discriminante alors qu'il faut près de 0.75 sec à SVM pour effectuer la même tâche.

Données Iris de Fisher Le tableau 6.3 présente les résultats de classification obtenus sur le jeu de données Iris de Fisher. Nous avons choisi de classifier également les données classiques que sont les données Iris de Fisher afin que chacun puisse comparer les résultats avec ceux d'autres méthodes. Nous avons été étonné d'obtenir de si bon résultats sur des données dont la dimension n'est pas si grande. Il semble que le fait de travailler dans des espaces différents pour chaque classe ait permis de séparer des données qui ne sont pas linéairement séparables. La figure 6.3 permet de visualiser les erreurs de classification et l'on peut voir que le seul point mal classé par l'HDDA l'est aussi par les autres méthodes.

Données LIS Le tableau 6.4 présente les résultats de classification obtenus sur le jeu de données LIS de catégorisation d'images. L'expérience menée sur les données LIS montre que la classification obtenue grâce à l'HDDA est meilleure que les autres méthodes bien que celles-ci réalisent des classifications correctes sur ce jeu de données. On remarque sur cet exemple que les méthodes classiques d'Analyse Discriminante sont fortement pénalisées par la grande dimension des données alors que l'HDDA et SVM ne subissent pas cet effet. De nouveau, l'HDDA réalise la classification des données aussi rapidement que les méthodes classiques d'Analyse Discriminante et beaucoup plus rapidement que SVM. Cet aspect peut être particulièrement intéressant dans le cadre de la reconnaissance de classes en vision, car cela ouvre la porte à la reconnaissance en temps réel.

7 Application à la reconnaissance de classes d'objets

La reconnaissance d'objets dans des images naturelles est un des problèmes les plus difficiles en vision par ordinateur. Ces dernières années, de nombreuses approches ont utilisé avec succès des descripteurs locaux d'images. Cependant, ces descripteurs locaux sont en grande dimension ce qui pénalise les méthodes de classification et par conséquent la reconnaissance. Pour cette raison, l'HDDA semble bien adaptée à cette application. Les méthodes généralement utilisées pour cette application sont l'Analyse Discriminante Linéaire (LDA) et, plus récemment, les mixtures de composantes principales [12] et les méthodes à noyaux [14, chap. 12]. De nombreuses études ont combiné une étape de réduction de dimension avec un classifieur classique : on peut citer les méthodes bien connues *Eigenfaces* [26] et *Fisherfaces* [2] qui réduisent la dimension des données respectivement par ACP et par projection sur les $(k-1)$ axes discriminants (FDA). Nous avons donc voulu vérifier si le classifieur HDDA surpassait le classifieur à noyaux SVM qui est actuellement le plus utilisé pour la reconnaissance d'objets.



Descripteurs détectés | Classifi cation par HDDA | Éléments de la moto | Probabilité d'erreur < 10^{-10}

FIG. 7.1 – Reconnaissance de la classe « moto » dans une image naturelle avec le classifieur HDDA.

7.1 La reconnaissance de classes d'objets

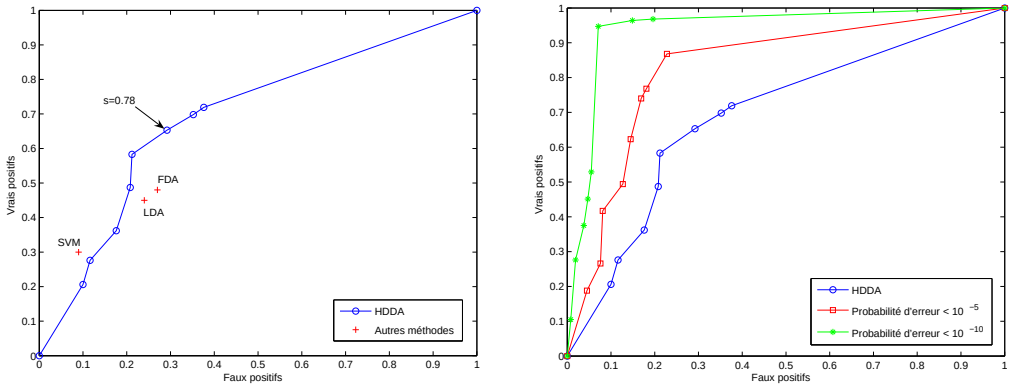
Le processus classique de reconnaissance d'objets se compose comme suit : tout d'abord, de petites régions de l'image sont détectées dans un ensemble d'apprentissage grâce au filtre de Harris-Laplace [19] puis sont décrites en utilisant un descripteur local invariant. Un objet est reconnu dans une image test si un nombre suffisant de correspondances avec le jeu d'apprentissage a été trouvé. Une récente étude [20] a montré que le descripteur SIFT [18] était particulièrement robuste aux variations d'échelle et de luminosité. Cependant, cette méthode fournit des données en haute-dimension, typiquement $d = 128$, ce qui pénalise la phase de décision. Des travaux antérieurs [5, 15] ont montré que le fait de réduire la dimension de ce type de données permet d'accroître le taux de reconnaissance. Pour cette raison, la méthode que nous proposons semble être appropriée. La figure 7.1 présente l'étape de classification de la reconnaissance de classe d'objet avec le classifieur HDDA : tous les descripteurs détectés dans l'image sont affectés à une des classes, puis on ne conserve que ceux appartenant aux sous-classes de l'objet et enfin ceux ayant une probabilité d'erreur de classement très faible. Nous présentons au paragraphe suivant les résultats obtenus avec notre méthode et nous les comparons à des méthodes classiques utilisées en reconnaissance de classe d'objet.

7.2 Les données

Pour cette application, nous avons choisi de travailler avec des images de motos (jeu de données disponible à l'adresse *http ://www.robots.ox.ac.uk/~vgg/*). Nous avons calculé les descripteurs pour un jeu de 200 images, puis nous avons sélectionné ceux correspondant à 3 parties caractéristiques de la moto : le guidon, la selle et les roues. Nous avons également retenu un certain nombre de descripteurs appartenant au fond afin de modéliser également cette classe. Nous avons ainsi obtenu un jeu de données comportant 2000 descripteurs en dimension 128 répartis en 4 classes : éléments de la selle, du guidon, des roues et éléments du fond.
Afin d'obtenir des résultats numériques, nous avons divisé le jeu de données précédent en un jeu d'apprentissage comportant 1500 descripteurs et un jeu de test comportant 500 descripteurs. Les proportions de chacune des classes valent respectivement $\pi_i = \frac{1}{5}$, $\forall i = 1, ..., 3$ et $\pi_4 = \frac{2}{5}$. Pour simuler une expérience de vision par ordinateur, nous avons également calculé les descripteurs d'images différentes de celles choisies pour l'apprentissage. Nous avons ensuite affecté chacun des descripteurs de chaque image à une des quatre classes définies *a priori* grâce à notre méthode de discrimination. Afin de comparer les performances de notre méthode à une méthode couramment utilisée dans ce type d'expérience, nous avons également classé les descripteurs avec la méthode SVM.

7.3 Résultats de classification

La figure 7.2-a présente les résultats de classification obtenus avec l'HDDA (modèle $[a_i b_i Q_i d_i]$), SVM (noyau gaussien), LDA et FDA sur ces données. Afin de synthétiser les résultats, seulement deux classes ont été considérées pour le tracé des courbes : moto (positif) et fond (négatif). On peut observer que la méthodes HDDA fournit des résultats bien meilleurs que ceux de la LDA et de la FDA. En effet, le fait



(a) (b)

FIG. 7.2 – (a) Comparaison des résultats de classification obtenus avec les méthodes HDDA, SVM, LDA et FDA. Les résultats de l'HDDA ont été obtenus pour différentes valeurs du seuil s. (b) Influence sur les résultats de classification obtenus avec le classifieur HDDA du seuillage sur la probabilité d'erreur.

que les résultats de la LDA et de la FDA soient en dessous de la courbe de l'HDDA signifie que l'on peut toujours trouver un seuil s tel que le résultat de l'HDDA soit meilleur que ceux de la LDA et de la FDA. On peut également remarquer que SVM semble surpasser notre méthode. Cependant, l'HDDA fournit une probabilité d'erreur de classement pour chaque descripteur (voir paragraphe 3.3) ce qui permet d'enlever le point dont la classification est douteuse. En effet, pour reconnaître un objet dans une image, il suffit d'identifier avec certitude quelques instances des parties de l'objet. La figure 7.2-b montre que si l'on ne conserve que les descripteurs dont la probabilité d'erreur est inférieure à un certain seuil, alors le taux de vrais positifs augmente significativement pour un taux de faux positifs fixé. Par exemple, si l'on ne conserve que les descripteurs ayant une probabilité d'erreur $< 10^{-10}$ (ce qui représente 15% des descripteurs), alors le taux de vrais positifs avoisine 90 % pour un taux de faux positifs inférieur à 10%. Le fondement probabiliste permet donc d'améliorer la reconnaissance en seuillant sur la probabilité d'erreur. De plus, on peut noter que l'HDDA est aussi rapide que les méthodes FDA et LDA (~ 1 sec.) et bien plus rapide que SVM (~ 7 sec.).

7.4 Résultats de reconnaissance

La figure 7.3 présente le résultat de la reconnaissance de la classe « moto » avec les classifieurs HDDA (modèle $[a_i b_i Q_i d_i]$ avec $s = 0.78$) et SVM (noyau gaussien) sur 10 images de motos différentes de celles du jeu d'apprentissage. Ces résultats montrent que le classifieur HDDA combiné au seuillage sur la probabilité d'erreur donne de meilleurs résultats que le classifieur SVM. En effet, les erreurs de classification sont significativement moins nombreuses avec HDDA qu'en utilisant SVM. Par exemple, si l'on considère la $5^{\text{ème}}$ image, HDDA reconnaît la moto sans erreurs alors que SVM commet 5 erreurs.

7.5 Perspectives

Ces premiers résultats prometteurs de notre méthode nous encouragent à continuer dans ce domaine d'application. De plus, notre méthode effectuant sa tâche de classification en un temps relativement court, cela ouvre la voie à la reconnaissance de classe d'objets en temps réel. En revanche, certaines contraintes existent à l'utilisation de notre méthode dans une application réelle. En particulier, la phase d'apprentissage requiert l'intervention humaine pour la sélection des parties caractéristiques de l'objet. Il serait donc profitable d'essayer de travailler dans un cadre le moins supervisé possible afin de rendre la tâche de reconnaissance la plus automatique possible. Pour cela, il nous faudrait adapter notre méthode à la classification

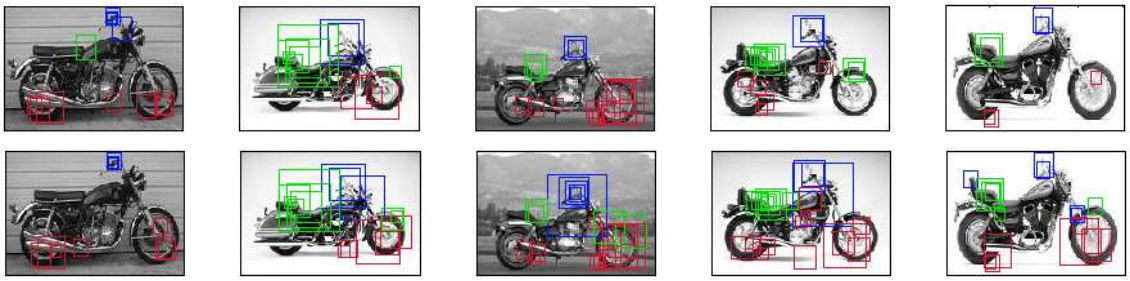


FIG. 7.3 – Reconnaissance de la classe « moto » avec les classifieurs HDDA (en haut) et SVM (en bas). Uniquement les descripteurs reconnu comme « moto » ont été affichés. Pour l'HDDA, seulement les descripteurs ayant une probabilité d'erreur inférieure à 10^{-10} ont été conservés. Les couleurs bleu, rouge et vert sont respectivement associées au guidon, aux roues et à la selle.

non supervisée. Nous pourrions utiliser le modèle statistique adapté aux données de grande dimension, présenté dans ce rapport, dans la méthode de classification automatique à modèle de mélanges.

8 Conclusion

Nous avons présenté dans ce rapport une nouvelle méthode de discrimination multi-classes, appelée Analyse Discriminante de Haute Dimension (*High Dimensionality Discriminant Analysis*), adaptée aux données de grande dimension. Cette méthode est basée sur l'Analyse Discriminante classique dont le modèle statistique a été adapté pour prendre en compte les spécificités des données de grande dimension. Nous proposons donc une nouvelle règle statistique de décision qui fournit, outre la classe d'un nouvel individu, une probabilité d'erreur qui pourra être utilisée comme indicateur de l'incertitude de classification. De plus, l'HDDA ne nécessite pas de pré-traitement des données, en particulier, il n'est pas nécessaire de réduire préalablement la dimension des données. Nous nous sommes également intéressé aux règles de décision induites pour des valeurs particulières des paramètres ; en particulier, nous montrons que l'Analyse Discriminante classique peut être vue comme un cas particulier de l'HDDA sous certaines hypothèses. Enfin, nous avons mis en œuvre et comparé l'HDDA à des méthodes de classification de référence sur des données artificielles et réelles. Nous appliquons notamment notre méthode à des données issues d'expériences de reconnaissance d'images où la rapidité de calcul et la donnée de la probabilité d'erreur sont des arguments prometteurs.

La méthode présentée dans ce rapport est une nouvelle voie pour l'analyse des données de grande dimension. Cependant, certains points techniques de la méthode méritent une étude plus approfondie. En particulier, l'estimation de la dimension intrinsèque des espaces propres de chaque classe pourrait être conduite par des méthodes fractales [6]. En outre, 10 cas particuliers restent à traiter : il faudrait calculer, pour ces modèles, les estimateurs dont les formulations ne sont pas explicites puis les implanter. D'autre part, notre méthode ne fournissant pas de projection des données, il serait bon de réfléchir à un moyen de visualiser le résultat de la classification. Enfin, une autre voie d'amélioration de l'HDDA a envisager est celle des récentes méthodes à noyaux [1, 24] que l'on pourrait adapter à notre méthode. Le modèle statistique des données de grande dimension que nous avons proposé dans ce rapport pourrait très certainement être combiné avec succès avec la méthode EM pour obtenir une classification automatique des données. C'est une des voies que nous allons développer pour l'appliquer à la reconnaissance de classes d'objets.

Références

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10) :2385–2404, 2000.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. 19(9) :711–720, 1997.

[3] R. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.

[4] H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91 :1743–1748, 1996.

[5] C. Bouveyron, S. Girard, and C. Schmid. Dimension reduction and classification methods for object recognition. In *5th French-Danish Workshop on Spatial Statistics and Image Analysis in Biology*, pages 109–113, May 2004.

[6] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10) :1404–1407, 2002.

[7] G. Celeux. Analyse discriminante. In G. Govaert, editor, *Analyse de Données*, pages 201–233. Hermes Science, Paris, France, 2003.

[8] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, Institut National Polytechnique de Grenoble, 1992.

[9] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7(2) :179–188, 1936.

[10] B. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79 :892–897, 1984.

[11] B. Flury and W. Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1) :169–184, 1984.

[12] B. Frey, A. Colmenarez, and T. Huang. Mixtures of local linear subspaces for face recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 32–37, 1998.

[13] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84 :165–175, 1989.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[15] Y. Ke and R. Sukthankar. Pca-sift : A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[16] H. Le Borgne. *Analyse de scènes naturelles par composantes indépendantes*. PhD thesis, Institut National Polytechnique de Grenoble, 2004.

[17] H. Le Borgne, N. Guyader, A. Guerin-Dugué, and J. Hérault. Classification of images : Ica filters vs human perception. In *7th International Symposium on Signal Processing and its Applications*, number 2, pages 251–254, 2003.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[19] K. Mikolajczk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, pages 525–531, 2003.

[20] K. Mikolajczk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[21] A. Mkhadri, G. Celeux, and A. Nasrollah. Regularization in discriminant analysis : an overview. *Computational Statistics and Data Analysis*, 23 :403–423, 1997.

[22] I. Pima and M. Aladjem. Regularized discriminant analysis for face recognition. *Pattern Recognition*, 37(9) :1945–1948, September 2004.

[23] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, Paris, France, 1990.

[24] B. Schölkopf, A. Smola, and K-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5) :1299–1319, 1998.

[25] D. Scott and J. Thompson. Probability density estimation in higher dimensions. In *Proceedings of the Fifteenth Symposium on the Interface, North Holland-Elsevier Science Publishers*, pages 173–179, 1983.

[26] M. Turk and A. Pentland. Eigenfaces for recognition. *Cognitive Neuroscience*, 3(1), 1991.

[27] M. Verleysen. Learning high-dimensional data. In S. Ablameyko, L. Goras, M. Gori, and V. Piuri, editors, *Limitations and future trends in neural computation*, pages 141–162. IOS Press, 2003.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399