Classification supervisée et non supervisée des données de grande dimension

Charles BOUVEYRON[1] & Stéphane GIRARD[2]

[1] SAMOS-MATISSE, CES, UMR CNRS 8174
Université Paris 1 (Panthéon-Sorbonne)
90 rue de Tolbiac, 75634 Paris Cedex 13, France

[2] Mistis, INRIA Rhône-Alpes & LJK
655 avenue de l'Europe, 38330 Saint-Ismier Cedex, France

Résumé Cet article est consacré à la classification des données de grande dimension. Supposant que de telles données vivent dans des sous-espaces de dimensions intrinsèques inférieures à la dimension de l'espace original, nous proposons une re-paramétrisation du modèle de mélange gaussien. En forçant certains paramètres à être communs dans une même classe ou entre les classes, nous exhibons une famille de modèles adaptés aux données de grande dimension, allant du modèle le plus général au plus parcimonieux. Ces modèles gaussiens sont ensuite utilisés pour la classification supervisée ou non-supervisée. La nature de notre re-paramétrisation permet aux méthodes ainsi construites de ne pas être perturbées par le mauvais conditionnement ou la singularité des matrices de covariance empiriques des classes et d'être efficaces en terme de temps de calcul.

Mots-clefs : Classification supervisée et non supervisée, fléau de la dimension, modèle de mélange gaussien, modèle parcimonieux.

1 Introduction

La classification de données situées dans un espace de grande dimension est un problème délicat qui apparaît dans de nombreuses sciences telles que l'analyse d'images. Dans cet article, nous focalisons notre attention sur les modèles probabilistes [12]. Parmi ceux-ci, le modèle de mélange gaussien est le plus populaire [33] bien que son comportement dans la pratique soit décevant lorsque la taille de l'échantillon est faible en regard du nombre de paramètres à estimer. Ce phénomène bien connu est appelé « fléau de la dimension » ou *curse of dimensionality* depuis les travaux de Bellman [3]. On pourra consulter [36, 37] pour une étude théorique de l'effet de la dimension en classification supervisée (ou discrimination).

Pour éviter le sur-ajustement des modèles, il est nécessaire de trouver un compromis entre le nombre de paramètres à estimer et la généricité du modèle. Nous proposons ici un modèle de mélange gaussien parcimonieux permettant de représenter le sous-espace propre à chacune des classes. Les paramètres de ce modèle sont estimés par maximum de vraisemblance ou par l'algorithme EM [20] selon que l'on soit confronté à un problème de classification supervisée ou non-supervisée. Les méthodes de classification ainsi construites sont baptisées respectivement HDDA pour *High Dimensional Discriminant Analysis* et HDDC pour *High Dimensional Data Clustering*. Notons qu'il est possible de contraindre le modèle afin de limiter davantage le nombre de paramètres à estimer. Ainsi, il sera

possible de supposer que les classes sont sphériques dans leur sous-espace propre ou de supposer que certaines de leurs caractéristiques sont communes à toutes les classes. La nature de notre modèle donne lieu à des méthodes HDDA et HDDC robustes aux mauvais conditionnements ou aux singularités éventuelles des matrices de covariance empiriques. Elles sont de plus très performantes en termes de temps de calcul.

La suite de cet article est organisée en quatre parties. Nous présentons dans le paragraphe 2 un état de l'art lié à la classification en grande dimension. Notre modèle de mélange gaussien est décrit au paragraphe 3. L'estimation de ses paramètres et donc la mise en œuvre des méthodes HDDA et HDDC fait l'objet du paragraphe 4. Enfin, quelques résultats expérimentaux obtenus sur données réelles et simulées sont présentés paragraphe 5.

2 Etat de l'art

Les méthodes classiques pour s'affranchir du fléau de la dimension consistent à réduire la dimension des données et/ou à utiliser un modèle de mélange gaussien parcimonieux. Quelques méthodes basées sur des modèles par sous-espaces ont également été introduites plus récemment. Nous proposons ci-dessous un tour d'horizon de ces trois familles de méthodes.

2.1 Réduction de dimension

De nombreuses méthodes utilisent une réduction de dimension globale pour s'affranchir du fléau de la dimension. Les plus simples d'entre elles consistent à réduire la dimension préalablement à une classification classique. A ce titre, l'analyse en composantes principales (ACP) [29] est très souvent utilisée en analyse d'images. L'ACP ne prenant en compte que les dépendances linéaires entre les variables, de nombreuses alternatives ont été proposées telles que l'ACP par noyaux [42], ACP par variétés [26, 28], ou réseaux de neuronnes [19, 30, 40, 46]. Cependant, déconnecter les phases de réduction de dimension et de classification ne semble pas judicieux. L'introduction d'une réduction de dimension dans l'analyse discriminante quadratique est étudiée dans [43]. La sélection de variables peut également être considérée comme une méthode de réduction de dimension. Il s'agit alors de choisir un sous-ensemble des variables représentatives des données, voir [27] pour une introduction plus complète. Une approche récemment introduite [39] consiste à combiner une sélection globale de variables dans le cadre d'un modèle de mélange gaussien. En règle générale, la réduction de dimension globale offre souvent de bonnes performances mais au prix d'une perte d'information qui aurait pu être discriminante. En effet, lorsque les classes sont localisées dans des sous-espaces différents, toute approche globale est inadaptée.

2.2 Modèles parcimonieux

Une solution altenative consiste à utiliser des modèles nécessitant l'estimation de peu de paramètres. Ainsi, il est possible de re-paramétrer les matrices de covariance des classes à partir de leur décomposition en éléments propres [2, 17] et, en contraignant certains paramètres à être commun à toutes les classes, on obtient alors des modèles parcimonieux.

D'autres modèles parcimonieux gaussiens sont introduits dans [25]. Ils forment une hiérarchie du plus complexe (une matrice de covariance pleine affectée à chaque groupe) au plus simple (une matrice de covariance identité commune à tous les groupes - modèle de l'algorithme des moyennes mobiles). Cependant, ces modèles ne peuvent rendre compte de l'existence d'un sous-espace propre spécifique à chaque classe.

2.3 Modélisation par sous-espaces

On distingue deux types de modélisation par sous-espaces. D'une part, les méthodes de poursuite de projection en classification [11, 18] supposent que les centres des classes sont situées dans un même sous-espace inconnu. A l'inverse, les méthodes de classification sur composantes principales (voir par exemple [10], Chapitre 17 ou [6]) reposent sur l'hypothèse que chaque classe est localisée dans un sous-espace qui lui est spécifique. Ainsi, l'analyse factorielle typologique [22] est basée sur un algorithme itératif semblable à celui des moyennes mobiles, alors que d'autres méthodes utilisent des techniques de recherche heuristiques [1]. Une synthèse de ce type de méthodes est dressée dans [35], la plupart d'entre elles étant fondées sur des considérations géométriques plus que sur des modèles probabilistes. La régression par classes (aussi connue sous le nom de *switching regression*) constitue une alternative intéressante et basée sur un modèle probabiliste. Les travaux [21, 38] en sont deux exemples, l'idée originale étant introduite dans [9]. Il a cependant été remarqué que la suppression de certaines directions est source d'instabilité en présence de données aberrantes ou de petits échantillons. Pour cette raison, les méthodes HDDA et HDDC que nous préconisons ne reposent pas sur l'élimination de directions prétendument inutiles mais modélisent les faibles variance par un unique paramètre. Les méthodes de mélanges d'analyses factorielles [34, 47] combinent un modèle à variables latentes avec un algorithme de type EM pour classer les données de grande dimension ou de dissimilarité [8]. Le nombre de paramètres du modèle de mélange gaussien est contrôlé par la dimension de la variable latente. Ce type de modèle permet de rendre compte des corrélations entre variables sans pour autant estimer des matrices de covariances pleines ou réduire leur dimension.

Le modèle introduit ci-dessous permet d'unifier certaines de ces approches par sous-espaces au sein d'un modèle de mélange gaussien tout en permettant l'introduction de contraintes de types modèles parcimonieux. Par ailleurs, une comparaison précise entre notre approche et les mélanges d'analyses factorielles est proposée dans [13].

3 Modèles gaussiens pour les grandes dimensions

La classification supervisée vise à associer chacune des n observations $\{x_1, ..., x_n\}$ à l'une des k classes connues *a priori* tandis que la classification non supervisée a pour but de regrouper ces données en k groupes homogènes. Le lecteur pourra trouver de plus amples détails sur ces deux approches dans [31] et [32]. L'approche la plus populaire dans ces deux situations est celle du modèle de mélange gaussien qui fait l'hypothèse que chaque classe est représentée par une densité de probabilité gaussienne. Cette approche suppose que les observations $\{x_1, ..., x_n\}$ sont des réalisations indépendantes d'un vecteur

aléatoire X à valeurs dans $\mathbb{R}^p$ de densité :

$$f(x,\theta) = \sum_{i=1}^{k} \pi_i \phi(x,\theta_i), \tag{1}$$

où ϕ est la densité de la loi normale multivariée de paramètres $\theta_i = \{\mu_i, \Sigma_i\}$ et π_i est la probabilité *a priori* de la $i$ème classe. Un tel modèle requiert l'estimation de matrices de covariance pleines et cela implique que le nombre de paramètres à estimer croît avec le carré de la dimension p. Cependant, le phénomène de « l'espace vide » [45] nous permet de conjecturer que la classification est une tâche plus aisée à réaliser dans des espaces de grande dimension. Nous allons par conséquent proposer une paramétrisation du modèle de mélange gaussien qui permette d'exploiter cette caractéristique des espaces de grande dimension. Notre idée est d'utiliser le fait que les données de grande dimension vivent dans des sous-espaces dont les dimensions intrinsèques sont faibles pour limiter le nombre de paramètres du modèle et régulariser l'estimation des matrices de covariance des classes.

3.1 Le modèle $[a_{ij}b_iQ_id_i]$

Nous nous plaçons dans le cadre classique du modèle de mélange gaussien et nous supposons que les densités conditionnelles des k classes sont gaussiennes $\mathcal{N}_p(\mu_i, \Sigma_i)$ de moyennes μ_i et de matrices de covariance Σ_i, pour $i = 1, ..., k$. Soit Q_i la matrice orthogonale composée des vecteurs propres de Σ_i, alors la matrice de covariance Δ_i est définie de la manière suivante dans l'espace propre de Σ_i :

$$\Delta_i = Q_i^t \, \Sigma_i \, Q_i. \tag{2}$$

La matrice Δ_i est par construction une matrice diagonale contenant les valeurs propres de Σ_i. Nous supposons en outre que Δ_i n'a que $d_i + 1$ valeurs propres différentes et a donc la forme suivante :

$$\Delta_i = \begin{pmatrix} \boxed{\begin{matrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_i & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b_i \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_i \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (p-d_i) \end{matrix} \tag{3}$$

avec $a_{ij} > b_i$, $j = 1, ..., d_i$, et où $d_i \in \{1, \ldots, p-1\}$ est inconnu. Le sous-espace spécifique $\mathbb{E}_i$ de la $i$ème classe est défini comme étant l'espace affine engendré par les d_i vecteurs propres associés aux valeurs propres a_{ij} et tel que $\mu_i \in \mathbb{E}_i$. De manière similaire, le sous-espace affine $\mathbb{E}_i^\perp$ est tel que $\mathbb{E}_i \oplus \mathbb{E}_i^\perp = \mathbb{R}^p$ et $\mu_i \in \mathbb{E}_i^\perp$. Dans le sous-espace $\mathbb{E}_i^\perp$, la variance est donc modélisée par l'unique paramètre b_i. Nous définissons également $P_i(x) = \tilde{Q}_i \tilde{Q}_i^{\,t}(x-\mu_i)+\mu_i$ et $P_i^\perp(x) = \bar{Q}_i \bar{Q}_i^t (x-\mu_i) + \mu_i$ les projections respectives de x sur $\mathbb{E}_i$ et $\mathbb{E}_i^\perp$, où $\tilde{Q}_i$ est composée des d_i premières colonnes de Q_i complétées par $(p-d_i)$ colonnes de zéros et $\bar{Q}_i = (Q_i - \tilde{Q}_i)$. Ainsi, la dimension d_i du sous-espace $\mathbb{E}_i$ peut être considérée comme la dimension intrinsèque de la $i$ème classe, *i.e.* le nombre de dimensions nécessaires pour une description satisfaisante de la $i$ème classe. La Figure 1 résume ces notations. En suivant le système de notation de [17], le modèle gaussien présenté dans ce paragraphe sera noté $[a_{ij}b_iQ_id_i]$ dans la suite.

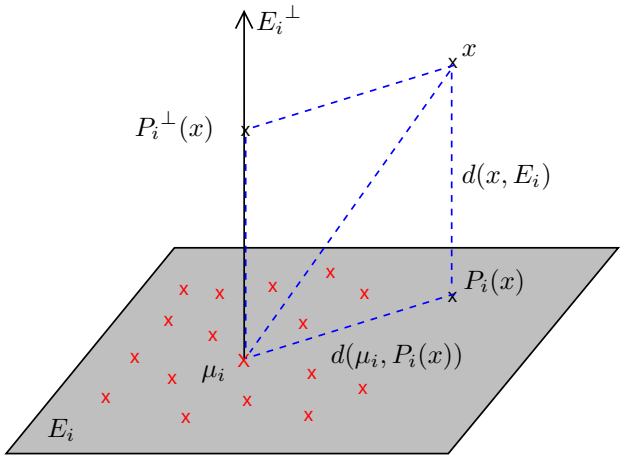

FIG. 1 – Les sous-espaces $\mathbb{E}_i$ et $\mathbb{E}_i^{\perp}$ de la $i$ème composante du mélange gaussien.

3.2 Le modèle $[a_{ij}b_iQ_id_i]$ et ses paramètres

La paramétrisation introduite ci-dessus permet de contrôler les caractéristiques de la $i$ème composante du mélange gaussien grâce à quatre types de paramètres : le vecteur $(a_{i1}, ..., a_{id_i})$, le scalaire b_i, la matrice Q_i et la dimension d_i. Les paramètres $a_{i1}, ..., a_{id_i}$ et b_i, contenus dans la matrice diagonale Δ_i, contrôlent la forme de la classe C_i. Plus particulièrement, les d_i valeurs $a_{i1}, ..., a_{id_i}$ paramètrent la forme de la densité dans le sous-espace $\mathbb{E}_i$ où vivent les données de la classe. Ces d_i paramètres représentent donc la dispersion réelle des données de la $i$ème classe. Le paramètre b_i modélise quant à lui la variance en dehors du sous-espace $\mathbb{E}_i$ qui est par conséquent supposée être isotropique. Ce paramètre représente donc la variance qui n'est pas due aux données de la classe et qui pourrait être due au bruit. La matrice orthogonale Q_i contrôle quant à elle l'orientation de la classe C_i par rapport au système des axes originaux. En particulier, les d_i premières colonnes de la matrice Q_i engendrent le sous-espace $\mathbb{E}_i$ où les données de la classe C_i sont sensées vivre. Enfin, le paramètre d_i, qui représente la dimension intrinsèque du sous-espace de la $i$ème classe, joue un rôle clé dans la paramétrisation que nous avons présentée. Nous verrons en effet au paragraphe 3.4 que c'est l'ensemble des paramètres d_i qui contrôle la complexité du modèle $[a_{ij}b_iQ_id_i]$.

3.3 Les sous-modèles du modèle $[a_{ij}b_iQ_id_i]$

En imposant certains paramètres à être communs entre les classes ou dans un même classe, nous obtenons des modèles particuliers qui correspondent à différentes régularisations du modèle $[a_{ij}b_iQ_id_i]$. Dans la suite, « Q_i libres » signifiera que chaque classe C_i a une matrice Q_i spécifique et « Q_i communes » traduira le fait que pour tout $i = 1, ..., k$, $Q_i = Q$ et donc que l'orientation des classes est la même. La famille du modèle $[a_{ij}b_iQ_id_i]$ compte 28 modèles et peut ainsi être divisée en trois catégories de modèles : les modèles à orientations libres (Q_i libres), les modèles à orientations communes (Q_i communes) et les modèles à matrices de covariance communes. La figure 2 permet d'observer l'influence de ces contraintes sur la forme des densités des classes. La représentation étant faite en

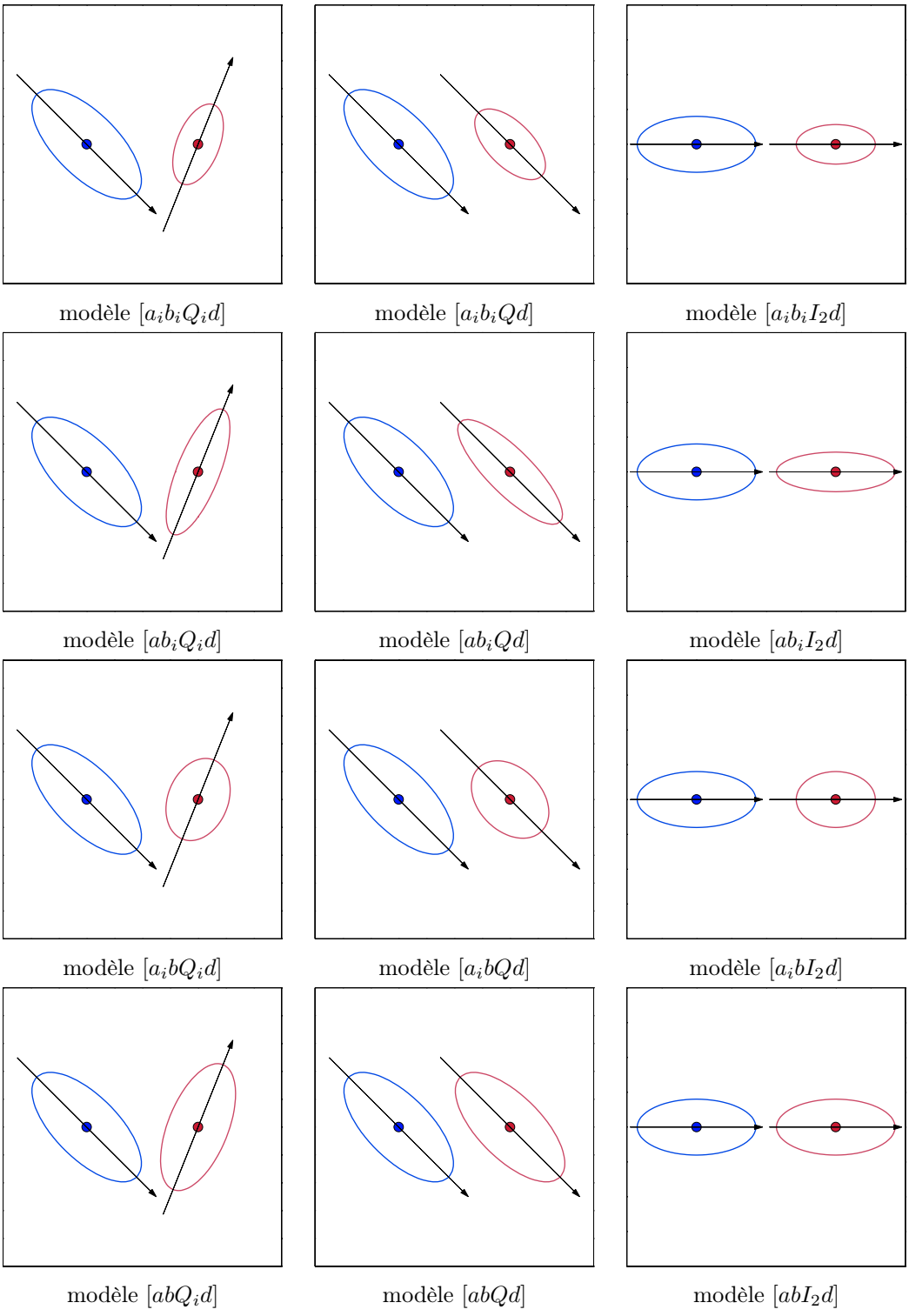

FIG. 2 – Influence des paramètres a_{ij}, b_i et Q_i sur les densités des classes. La représentation étant faite en dimension 2, les dimensions intrinsèques d_i des classes sont fixées et égales à 1.

dimension 2, les dimensions intrinsèques d_i des classes sont fixées et égales à 1. Les modèles de la première ligne font l'hypothèse que les variances dans et en dehors de leur sous-espace spécifique sont propres pour chacune des deux classes. A la seconde ligne, les modèles supposent que les variances dans les sous-espaces spécifiques sont égales. A l'inverse, les modèles de la troisième ligne supposent que la variance en dehors des sous-espaces des classes est commune. Enfin, la dernière ligne de la figure présente les modèles faisant l'hypothèse que les variances dans et en dehors des sous-espaces sont communes entre les classes.

Modèles à orientations libres Ces modèles supposent que les classes vivent dans des sous-espaces d'orientations différentes, *i.e.* les matrices Q_i sont spécifiques à chaque classe. Le modèle général $[a_{ij}b_iQ_id_i]$ appartient naturellement à cette catégorie. Notons tout d'abord qu'il est possible de supposer que $d_i = (p-1)$ pour tout $i = 1, ..., k$ et, dans ce cas, le modèle $[a_{ij}b_iQ_id_i]$ est en fait le modèle gaussien classique avec des matrices de covariances pleines pour chaque classe. Ce modèle donne naissance dans le cadre supervisé à la populaire méthode de l'Analyse Discriminante Quadratique (QDA), voir par exemple [41], paragraphe 18.5.1. D'autre part, en contraignant les dimensions d_i à être communes entre les classes, le modèle général donne naissance au modèle $[a_{ij}b_iQ_id]$ qui correspond au modèle proposé dans [47]. De ce fait, notre approche inclut le modèle de mélange d'analyses en composantes principales probabilistes (*Probabilistic Principal Component Analyzers*) introduit dans [47] et étendu dans [34]. Dans notre modélisation, d_i dépend de la classe et cela permet de modéliser une dépendance entre le nombre de facteurs principaux et les classes. De plus, notre approche peut être combinée avec une stratégie de modèles parcimonieux pour limiter encore le nombre de paramètres à estimer. En imposant aux d_i premières valeurs propres à être égales pour chaque classe, nous obtenons le modèle contraint $[a_ib_iQ_id_i]$. Nous avons noté que, en pratique, ce modèle donne souvent des résultats satisfaisants, *i.e.* l'hypothèse que chaque matrice Δ_i ne contient que deux valeurs propres différentes, a_i et b_i, semble être un moyen efficace de régulariser l'estimation de Δ_i. Un autre moyen de régularisation est de fixer les paramètres b_i à être communs entre les classes. Cette hypothèse donne naissance au modèle $[a_ibQ_id_i]$ qui suppose que la variance en dehors des sous-espaces spécifiques est commune. Cela peut être interprété comme la modélisation du bruit dans $\mathbb{E}_i^{\perp}$ par un unique paramètre b, ce qui est plutôt naturel si les données ont été acquises selon le même protocole. Cette catégorie contient également les modèles $[ab_iQ_id_i]$, $[abQ_id_i]$ et tous les modèles avec Q_i libre et d_i commun.

Modèles à orientations communes Il est également possible de supposer que l'orientation des classes est commune, *i.e.* $Q_i = Q$ pour tout $i = 1, ..., k$. Cependant, Il est important de remarquer que cette hypothèse n'implique pas que les sous-espaces spécifiques soient les mêmes. En effet, les moyennes des classes étant différentes pour chaque classe, les sous-espaces sont au plus parallèles. Cette hypothèse peut s'avérer intéressante pour modéliser des classes ayant des propriétés communes tout en gardant certaines spécificités. Plusieurs modèles de cette catégorie nécessitent l'utilisation de l'algorithme itératif FG [23] et ne seront donc pas considérés dans la suite de l'article. Par conséquent, seulement les modèles $[a_ib_iQd]$, $[ab_iQd]$ et $[a_ibQd]$ seront dorénavant considérés car leurs paramètres peuvent être estimés grâce à une procédure itérative simple. Remarquons au

passage qu'un modèle similaire au modèle $[a_{ij}bQd]$ a été considéré par Flury *et al.* [24] dans le contexte supervisé avec des hypothèses supplémentaires sur les moyennes.

Modèles à matrices de covariance commune Cette branche de la famille ne comporte que deux modèles : les modèles $[a_jbQd]$ et $[abQd]$. Ces deux modèles supposent en effet que les k classes ont même matrice de covariance $\Sigma = Q\Delta Q^t$. En particulier, si l'on fixe $d = (p-1)$, le modèle $[a_jbQd]$ revient au modèle gaussien, noté *Com-GMM* dans la suite, qui donne naissance dans le cadre supervisé à la méthode bien connue de l'Analyse Discriminante Linéaire (LDA, [41], paragraphe 18.5.1). Remarquons également que si $d < (p-1)$, le modèle $[a_jbQd]$ peut être vu comme une combinaison d'une méthode de réduction de dimension avec un modèle de mélange gaussien à matrices de covariance communes, mais cela sans perte d'information puisque l'information portée par les plus petites valeurs propres est conservée.

3.4 Complexité des différents modèles

La famille de modèles présentée dans les paragraphes précédents ne requiert que l'estimation de sous-espaces de dimension d_i et, de ce fait, les différents modèles de cette famille sont significativement plus parcimonieux que les modèles gaussiens classiques si $d_i \ll p$. Considérons en particulier le cas de données vivant dans un espace de dimension 100, composées de 4 classes de dimensions intrinsèques d_i égales à 10. Dans un tel cas, le modèle $[a_{ij}b_iQ_id_i]$ ne requiert l'estimation que de 4.231 paramètres là où le modèle gaussien plein et le modèle gaussien à matrices de covariance égales nécessitent respectivement l'estimation de 20.603 et 5.453 paramètres. En outre, le modèle $[a_{ij}b_iQ_id_i]$, qui donne naissance à une règle de décision quadratique, requiert l'estimation de moins de paramètres que le modèle gaussien à matrices de covariance égales qui donne, lui, une règle de décision linéaire.

4 Estimation des paramètres

Nous allons à présent considérer l'estimation des différents paramètres de modèles gaussiens pour la grandes dimension présentés précédemment. Nous traiterons les cas supervisé et non supervisé puisque ces modèles sont utilisables dans les deux contextes. Cependant, dans un souci de clarté, nous présenterons uniquement les estimateurs des modèles à orientations libres.

4.1 Le cas supervisé : la méthode HDDA

L'utilisation dans le cadre de la classification supervisée des modèles gaussiens pour la grande dimension a donné naissance à la méthode *High-Dimensional Discriminant Analysis* (HDDA) [14]. Dans ce contexte, les données d'apprentissage étant complètes, *i.e.* un label z indiquant la classe d'appartenance est associé à chaque observation x, l'estimation des paramètres du modèle par maximum de vraisemblance est directe et conduit aux estimateurs suivants. Les proportions du mélange ainsi que les moyennes

sont respectivement estimées par :

$$\hat{\pi}_i = \frac{n_i}{n},\ \hat{\mu}_i = \frac{1}{n_i} \sum_{j/z_j=i} x_j,$$

où n_i est le nombre d'individus dans la $i$ème classe et z_j indique le numéro de la classe de l'observation x_j. Nous introduisons de plus W_i, la matrice de covariance empirique de la $i$ème classe, définie par :

$$W_i = \frac{1}{n_i} \sum_{j/z_j=i} (x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)^t.$$

L'estimation des paramètres spécifiques du modèle introduit précédemment est détaillée ci-après. Le détail des calculs et les estimateurs pour les autres modèles sont donnés dans [14]. Les estimateurs du maximum de vraisemblance des paramètres des modèles à orientations libres sont explicites et donnés par :

- Matrice d'orientation Q_i : les d_i premières colonnes de Q_i sont estimées par les vecteurs propres associés aux d_i plus grandes valeurs propres λ_{ij} de W_i.
- Modèle $[a_{ij}b_iQ_id_i]$: l'estimateur de a_{ij} est $\hat{a}_{ij} = \lambda_{ij}$ et l'estimateur de b_i est la moyenne des $(p - d_i)$ plus petites valeurs propres de W_i. Il peut être écrit comme suit :

$$\hat{b}_i = \frac{1}{(p - d_i)} \left(\mathrm{tr}(W_i) - \sum_{j=1}^{d_i} \lambda_{ij} \right), \tag{4}$$

  où $\mathrm{tr}(W_i)$ est la trace de la matrice W_i.
- Modèle $[a_{ij}bQ_id_i]$: l'estimateur de a_{ij} est $\hat{a}_{ij} = \lambda_{ij}$ et l'estimateur de b est :

$$\hat{b} = \frac{1}{(p - \xi)} \left(\mathrm{tr}(W) - \sum_{i=1}^{k} \hat{\pi}_i \sum_{j=1}^{d_i} \lambda_{ij} \right), \tag{5}$$

  où $\xi = \sum_{i=1}^{k} \hat{\pi}_i d_i$ et $W = \sum_{i=1}^{k} \hat{\pi}_i W_i$ est la matrice de covariance intra-classe.
- Modèle $[a_ib_iQ_id_i]$: l'estimateur de b_i est donné par (4) et l'estimateur de a_i est :

$$\hat{a}_i = \frac{1}{d_i} \sum_{j=1}^{d_i} \lambda_{ij}. \tag{6}$$

- Modèle $[ab_iQ_id_i]$: l'estimateur de b_i est donné par (4) et l'estimateur de a est :

$$\hat{a} = \frac{1}{\xi} \sum_{i=1}^{k} \hat{\pi}_i \sum_{j=1}^{d_i} \lambda_{ij}. \tag{7}$$

- Modèle $[a_ibQ_id_i]$: les estimateurs de a_i et b sont respectivement donnés par (6) et (5).
- Modèle $[abQ_id_i]$: les estimateurs de a and b sont respectivement donnés par (7) et (5).

De façon classique, la classification d'un nouvelle observation $x \in \mathbb{R}^p$ se fait grâce à la règle du *maximum a posteriori* (MAP) qui affecte l'observation x à la classe la plus probable *a posteriori*. Ainsi, l'étape de classification consiste principalement à calculer $\mathbb{P}(Z = i|X = x)$ pour chaque classe $i = 1, ..., k$:

$$\mathbb{P}(Z = i|X = x) = 1 \Big/ \sum_{\ell=1}^{k} \exp\left(\frac{1}{2}(K_i(x_j) - K_\ell(x_j))\right),$$

où $K_i(x) = -2\log(\pi_i\phi(x, \theta_i))$ a la forme suivante dans le cas du modèle $[a_i b_i Q_i d_i]$:

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

Remarquons que $K_i(x)$ est principalement basée sur deux distances (illustrées Figure 1) : la distance entre la projection de x sur $\mathbb{E}_i$ et la moyenne de la classe et la distance entre l'observation et le sous-espace $\mathbb{E}_i$. Cette fonction de coût favorise l'affectation d'une nouvelle observation à la classe pour laquelle il est à la fois proche du sous-espace et pour laquelle sa projection sur le sous-espace de la classe est proche de la moyenne de la classe. Les termes de variance a_i et b_i pondèrent l'importance de ces deux distances. Par exemple, si les données sont très bruitées, *i.e.* b_i grand, il est naturel de pondérer la distance $\|x - P_i(x)\|^2$ par $1/b_i$ afin de tenir compte de la grande variance dans $\mathbb{E}_i^\perp$.

4.2 Le cas non supervisé : la méthode HDDC

L'utilisation dans le cadre de la classification non supervisée des modèles gaussiens pour la grande dimension a donné naissance à la méthode *High-Dimensional Data Clustering* (HDDC) [13]. Dans ce contexte, les données d'apprentissage n'étant pas complètes, *i.e.* le label z indiquant la classe d'appartenance est manquant pour chaque observation x, l'estimation des paramètres du modèle par maximum de vraisemblance n'est pas directe et nécessite l'utilisation d'un algorithme itératif : l'algorithme EM [20]. Le lecteur pourra consulter [32] pour plus de détails sur l'algorithme EM et ses extensions. En particulier, les modèles présentés dans cet article peuvent également être combinés avec les algorithmes *Classification EM* et *Stochastic EM* [16]. Avec les hypothèses et notations des modèles à orientations libres, l'algorithme EM prend la forme suivante :

Etape E : Cette étape calcule à l'itération q et pour chaque $i = 1, ..., k$ et $j = 1, ..., n$, la probabilité conditionnelle $t_{ij}^{(q)} = \mathbb{P}(x_j \in C_i^{(q-1)}|x_j)$ qui peut s'écrire à partir de (1) et en utilisant la règle de Bayes comme suit :

$$t_{ij}^{(q)} = 1 \Big/ \sum_{\ell=1}^{k} \exp\left(\frac{1}{2}(K_i^{(q-1)}(x_j) - K_\ell^{(q-1)}(x_j))\right),$$

avec $K_i^{(q-1)}(x) = -2\log(\pi_i^{(q-1)}\phi(x, \theta_i^{(q-1)}))$ et où $\pi_i^{(q-1)}$ et $\theta_i^{(q-1)}$ sont les paramètres du mélange estimés dans l'étape M à l'itération $(q - 1)$.

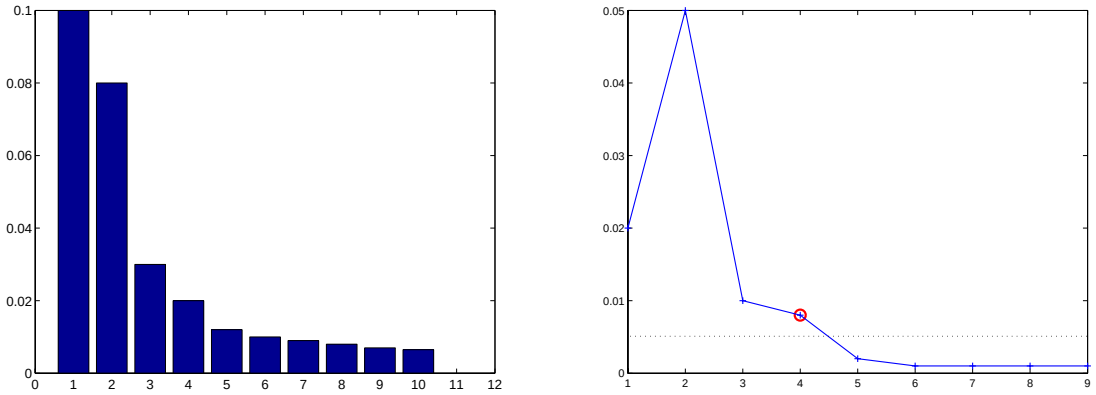

FIG. 3 – Estimation des dimensions intrinsèques d_i en utilisant le *scree-test* de Cattell : éboulis des valeurs propres de W_i (gauche) et différences entre les valeurs propres consécutives (droite). Les points sont reliés pour une plus grande lisibilité.

Etape M : Cette étape maximise à l'itération q la vraisemblance conditionnellement aux $t_{ij}^{(q)}$. Les estimateurs des proportions du mélange et des moyennes sont :

$$\hat{\pi}_i^{(q)} = \frac{n_i^{(q)}}{n}, \ \hat{\mu}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ij}^{(q)} x_j,$$

où $n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)}$. Nous introduisons de plus $W_i^{(q)}$, la matrice de covariance empirique du $i$ème groupe, définie par :

$$W_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ij}^{(q)} (x_j - \hat{\mu}_i^{(q)})(x_j - \hat{\mu}_i^{(q)})^t.$$

A l'itération q, les estimateurs des paramètres a_{ij}, b_i et Q_i, spécifiques aux modèles gaussiens pour les grandes dimensions, sont les mêmes que dans le cas supervisé.

4.3 Estimations des hyper-paramètres

L'estimation des paramètres, que ce soit dans le cadre supervisé ou non supervisé, requiert la connaissance de la dimension intrinsèque de chaque classe. L'estimation des dimensions intrinsèques est un problème difficile pour lequel il n'y a pas de solution universelle. L'approche que nous proposons est basée sur les valeurs propres de la matrice de covariance empirique W_i de chacune des classes. En effet, la $j$ème valeur propre de W_i correspond à la part de la variance totale portée par le $j$ème vecteur propre de W_i. Nous proposons d'estimer la dimension intrinsèque d_i, $i = 1, ..., k$ grâce au *scree-test* de Cattell [15] qui recherche un coude dans l'éboulis des valeurs propres. La dimension sélectionnée est la dimension pour laquelle les différences entre les valeurs propres sont plus petites qu'un seuil. La Figure 3 illustre le principe de la méthode. Dans cet exemple, quatre dimensions seront sélectionnées et cela correspond bien à un coude dans l'éboulis des valeurs propres. En pratique, nous recommandons de fixer le seuil à 0.2 fois la valeur de la plus grande différence. Dans le cas non supervisé, il est également nécessaire de déterminer le nombre k de composantes du mélange et cela peut être fait grâce au critère BIC [44].

4.4 Considérations numériques

Il est tout d'abord important de remarquer que la paramétrisation des modèles gaussiens présentés dans cet article fournit une expression explicite de Σ_i^{-1} alors que les méthodes classiques doivent inverser numériquement la matrice Σ_i et échouent généralement du fait de la singularité de la matrice. De plus, en observant l'expression de $K_i(x) = -2\log(\pi_i\phi(x,\theta_i))$, on remarque que le calcul des probabilités *a posteriori* n'utilise pas la projection sur $\mathbb{E}_i^{\perp}$ et par conséquent ne nécessite pas le calcul des $(p-d_i)$ dernières colonnes de la matrice d'orientation Q_i. Les méthodes HDDA et HDDC ne dépendent donc pas de la détermination de ces axes associés aux plus petites valeurs propres dont l'estimation en grande dimension est généralement instable. Ainsi, les méthodes de classification HDDA et HDDC ne sont pas perturbées par le mauvais conditionnement ou la singularité des matrices de covariance empiriques des classes. En outre, le fait de n'avoir qu'à déterminer les d_i plus grandes valeurs propres ainsi que leur vecteur propre associé se révèle être d'un intérêt crucial quand le nombre d'observations est plus petit que la dimension de l'espace original. Dans ce cas, il est préférable d'un point de vue numérique de calculer les valeurs propres et les vecteurs propres de la matrice $\Upsilon_i\Upsilon_i^t$ au lieu de calculer ceux de la matrice de covariance empirique $W_i = \Upsilon_i^t\Upsilon_i$, où Υ_i est la matrice $n_i \times p$ contenant les n_i observations centrées de la $i$ème classe vivant dans $\mathbb{R}^p$. En effet, la matrice W_i étant de dimension $p \times p$, la détermination des vecteurs propres associés aux d_i plus grandes valeurs propres de W_i est beaucoup plus longue et instable numériquement que la détermination des vecteurs propres associés aux d_i plus grandes valeurs propres de la matrice $\Upsilon_i\Upsilon_i^t$ si $n_i < p$. Le vecteur propre de la matrice $W_i = \Upsilon_i^t\Upsilon_i$ associé à la valeur propre λ_{ij} s'obtient à partir du vecteur propre v_{ij} de $\Upsilon_i\Upsilon_i^t$ associé à λ_{ij} en le multipliant à gauche par Υ_i^t. Typiquement, dans le cas de données où chacune des classes est représentée par 13 observations en dimension 1024 dans le jeu d'apprentissage, la détermination des vecteurs propres associés aux d_i plus grandes valeurs propres de $\Upsilon_i\Upsilon_i^t$ s'avère être 500 fois plus rapide que la détermination des vecteurs propres associés aux d_i plus grandes valeurs propres de W_i.

4.5 Les modèles gaussiens HD dans le logiciel MixMod

Parmi les modèles gaussiens pour la grande dimension (modèles HD ci-après), 8 modèles ont été sélectionnés pour être implantés dans le logiciel MixMod (disponible à l'adresse *http ://www-math.univ-fcomte.fr/mixmod/*) qui permet de traiter des problématiques d'estimation de densités, de classification ou d'analyse discriminante. Il propose notamment un large choix d'algorithmes d'estimation, dont EM et ses versions stochastique (SEM) ou de classification (CEM) afin de maximiser la vraisemblance ou la vraisemblance complétée. MIXMOD peut être utilisé sur des données quantitatives (grâce aux modèles de mélanges gaussiens multidimensionnels) ou qualitatives (grâce aux modèles de mélanges multinomiaux multidimensionnels). De plus, différents critères (BIC, ICL, NEC, CV) permettent de sélectionner le meilleur modèle parmi plusieurs modèles parcimonieux proposés : 22 modèles dans le cadre quantitatif (dont 8 sont spécifiques aux données de haute dimension) et 5 dans le cadre qualitatif. Ecrit en C++, MIXMOD est interfacé avec Scilab et Matlab et distribué sous la licence GNU (GPL). Les 8 modèles HD disponibles dans MixMod comportent :

– deux modèles à dimensions d_i libres :

  - le modèle $[a_{ij}b_iQ_id_i]$
  - le modèle $[a_ib_iQ_id_i]$
- six modèles à dimensions d_i communes :
  - le modèle $[a_{ij}b_iQ_id]$
  - le modèle $[a_jb_iQ_id]$
  - le modèle $[a_{ij}bQ_id]$
  - le modèle $[a_jbQ_id]$
  - le modèle $[a_ib_iQ_id]$
  - le modèle $[a_ibQ_id]$

Pour une première utilisation des modèles HD dans MixMod, nous recommandons l'utilisation de l'interface graphique. L'utilisateur intéressé par une utilisation plus avancée en ligne de commande dans Matlab ou Scilab pourra exécuter Mixmod avec les fichiers de configuration `HD_USPS_358_M.test` et `HD_USPS_358_MAP.test` disponibles dans le dossier `MIXMOD/TEST`.

5 Résultats expérimentaux et applications

Dans ce paragraphe, nous présentons une évaluation numérique des méthodes HDDA et HDDC, sur des jeux de données artificielles et réelles, illustrant les principales caractéristiques de ces deux méthodes. Dans la suite des expériences, HDDA et HDDC seront comparées à différents modèles classiques de mélange gaussien : modèle gaussien avec une matrice de covariance pleine pour chaque classe (QDA ou Full-GMM), avec une matrice de covariance commune pour toutes les classes (LDA ou Comm-GMM), avec des matrices de covariance diagonales (Diag-GMM) et avec des matrices de covariance sphériques (Sphe-GMM).

5.1 Influence de la dimension

Cette première expérience vise à mettre en lumière l'effet de la dimension sur différents modèles gaussiens dans le cadre de la classification supervisée. Pour cela, nous avons simulé trois classes modélisées par des densités gaussiennes dans $\mathbb{R}^p$, $p = 15, ..., 100$, selon le modèle $[a_ib_iQ_id_i]$ avec les paramètres suivants : $\{d_1, d_2, d_3\} = \{2, 5, 10\}$, $\{\pi_1, \pi_2, \pi_3\} = \{0.4, 0.3, 0.3\}$, $\{a_1, a_2, a_3\} = \{150, 75, 50\}$, $\{b_1, b_2, b_3\} = \{10, 10, 10\}$ et avec des moyennes relativement proches et des matrices d'orientations Q_i aléatoires. Les jeux de données d'apprentissage et de test sont respectivement composés de 250 et 1000 points. La performance des différentes méthodes étudiées a été mesurée par le taux moyen de classification correcte pour le jeu de test et calculée sur 50 répétitions de la simulation. La Figure 4 présente les taux moyens de classification correcte sur le jeu de test pour les méthodes HDDA (modèle gaussien $[a_ib_iQ_id_i]$), QDA, LDA, PCA+LDA, EDDA [4] et le classifieur optimal de Bayes (paramètres du modèle connus et non estimés). Nous pouvons tout d'abord observer que l'augmentation de la dimension des données n'affecte pas la performance de HDDA et que ses résultats sont très proches du classifieur optimal. De plus, HDDA fournit un taux de classification correcte similaire à celui de QDA en dimension faible et cela montre bien l'aspect quadratique de HDDA. La méthode QDA, qui est connue pour être particulièrement sensible à la dimension des données, donne en effet des résultats très décevant dès que la dimension augmente. La méthode LDA apparaît être moins sensible

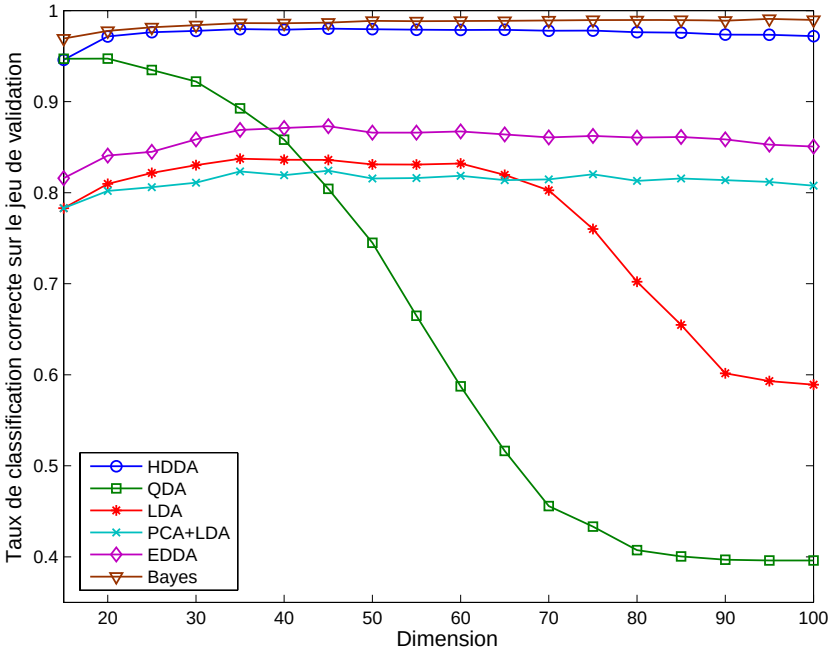

FIG. 4 – Influence de la dimension sur le taux de classification correcte sur un jeu de test et pour différents modèles gaussiens.

Modèle	Variables	Tx de classif. correcte
Sphe-GMM	Originales	0.605
VS-GMM	Originales	0.925
Sphe-GMM	Composantes principales	0.605
VS-GMM	Composantes principales	0.935
HDDC $[a_i b_i Q_i d_i]$	Originales	**0.950**

TAB. 1 – Taux de classification correcte pour le jeu de données « Crabes » pour différents modèles dans le cadre non supervisé.

à l'augmentation de la dimension mais fournit des résultats toujours décevants face à ces données complexes. Alors que LDA est fortement pénalisée pour des dimensions plus grandes que 60, une étape de réduction de dimension par ACP permet à PCA+LDA de fournir des résultats constants mais sans améliorer la performance initiale de LDA. Enfin, la méthode EDDA, et son modèle parcimonieux $[\lambda_k B_k]$ recommandé par les auteurs, ne semble pas souffrir du fléau de la dimension, mais fournit des résultats moins bons que ceux de l'HDDA. Pour résumer, cette expérience confirme bien que la méthode HDDA (modèle gaussien $[a_i b_i Q_i d_i]$) n'est pas sensible à la grande dimension des données et fournit de bons résultats aussi bien en dimension faible qu'en grande dimension.

5.2 Comparaison à la sélection de variables

Dans cette expérience, nous allons comparer l'utilisation de modèles gaussiens dans des sous-espaces à la sélection de variables dans le cadre de la classification non supervisée.

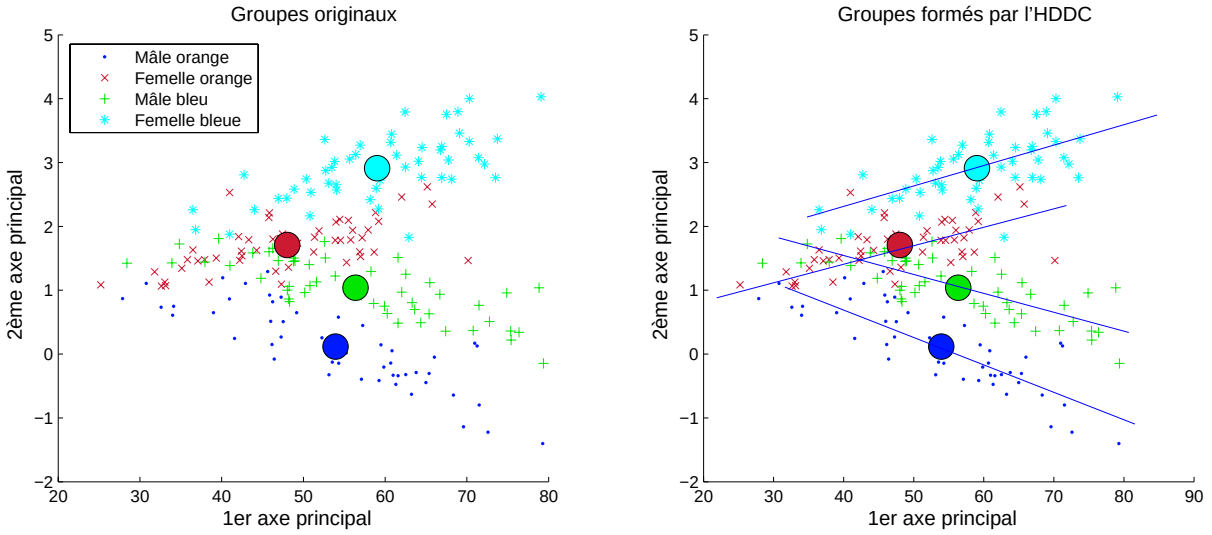

FIG. 5 – Classification non supervisée grâce à l'HDDC des données « Crabes » : à gauche, données projetées sur le premier plan principal et, à droite, segmentation obtenue avec le modèle $[a_i b_i Q_i d_i]$ de l'HDDC ainsi que les sous-espaces spécifiques estimés (lignes bleues).

Les données « crabes », utilisées dans cette mise en pratique de l'algorithme de type EM qu'est l'HDDC, sont composées de 5 mesures faites sur 200 individus équi-répartis dans 4 classes : les crabes mâles et femelles à carapace orange, les crabes mâles et femelles à carapace bleue. Pour chacun des sujets, 5 variables ont été observées : largeur de la lèvre frontale, largeur arrière, longueur de la carapace, largeur maximale de la carapace et profondeur du corps de l'animal. La Figure 5 présente les données sur le premier plan principal. Les données que nous considérons ici présentent l'intérêt que les dimensions intrinsèques des sous-espaces spécifiques des groupes sont égales à 1. Cette spécificité des données permettra donc une visualisation aisée de la recherche des sous-espaces des classes par l'HDDC. La figure 6 montre les 12 étapes de l'algorithme d'estimation des paramètres du modèle $[a_{ij} b_i Q_i d_i]$ sur les données « crabes ». Les données sont projetées sur les deux premiers axes principaux pour la visualisation uniquement. Nous rappelons que l'HDDC, comme l'HDDA, ne réduit jamais la dimension des données mais le modèle gaussien sous-jacent tient compte du fait que la dimension intrinsèque des données de chaque classe est plus petite que p. Les sous-espaces spécifiques des composantes du mélange sont représentés sur la figure par des lignes bleues et les moyennes des classes floues sont symbolisées par des disques de couleurs.

Nous nous proposons ici de comparer notre approche à la sélection de variables. Une récente approche, appelée VS-GMM dans la suite, proposée par Raftery et Dean [39] permet de combiner la sélection de variables à l'étape de classification dans le cadre du modèle de mélange gaussien. Pour ce faire, les auteurs considèrent le problème de la sélection de variables comme un problème de choix de modèles. Le tableau 1 présente les résultats de classification obtenus avec l'HDDC et les méthodes usuelles de classification non supervisée. Nous y avons également reporté le résultat de VS-GMM donné par [39]. Il apparaît tout d'abord que ces données, qui ne sont certes pas de grande dimension, s'avèrent très difficiles à classer avec les méthodes classiques. En effet, le meilleur résultat obtenu avec une méthode classique est 0.64 en utilisant le modèle gaussien général full-GMM. On re-

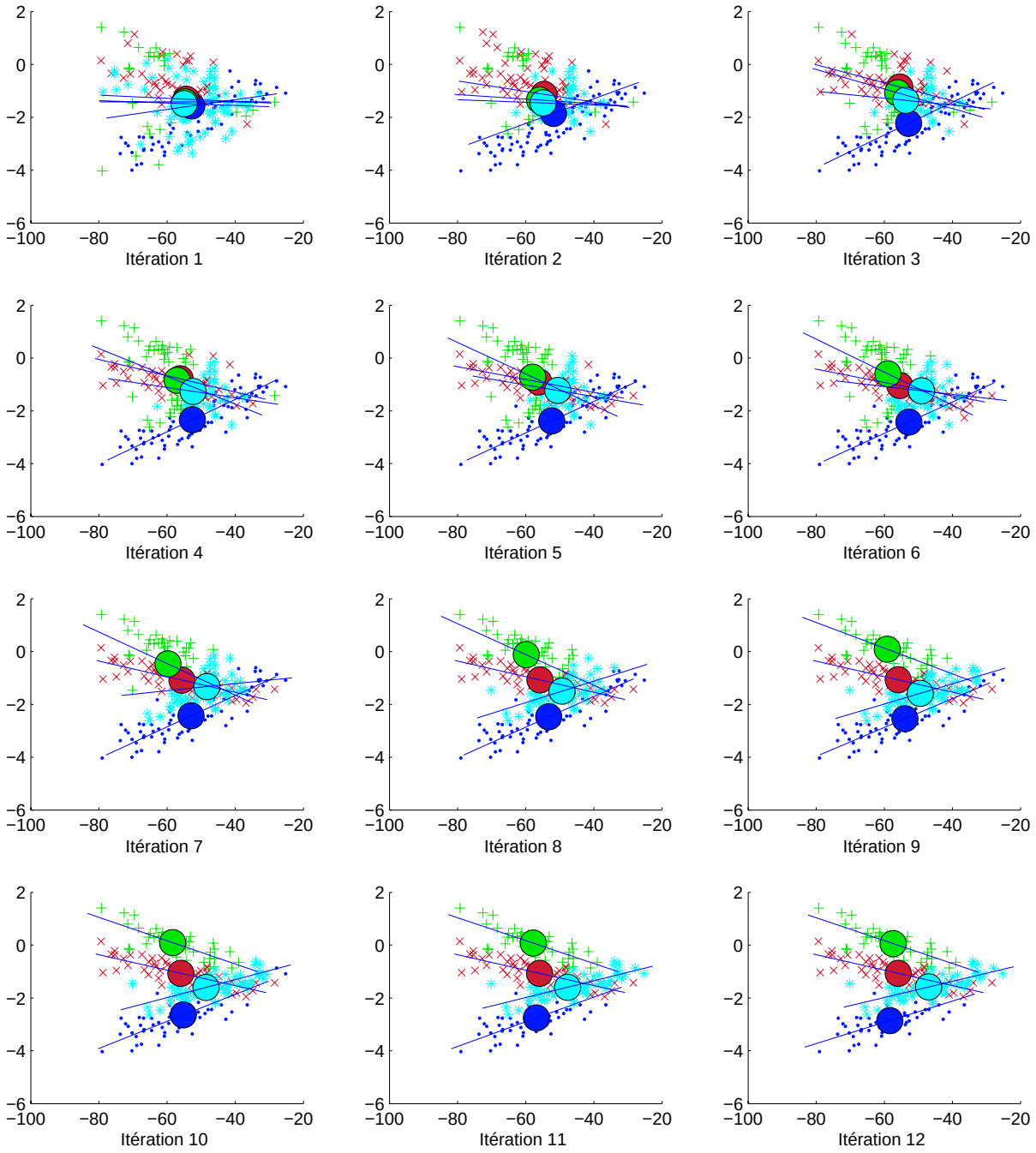

FIG. 6 – Les étapes de l'algorithme EM de l'HDDC sur les données « Crabes » ainsi que les sous-espaces spécifiques estimés (lignes bleues).

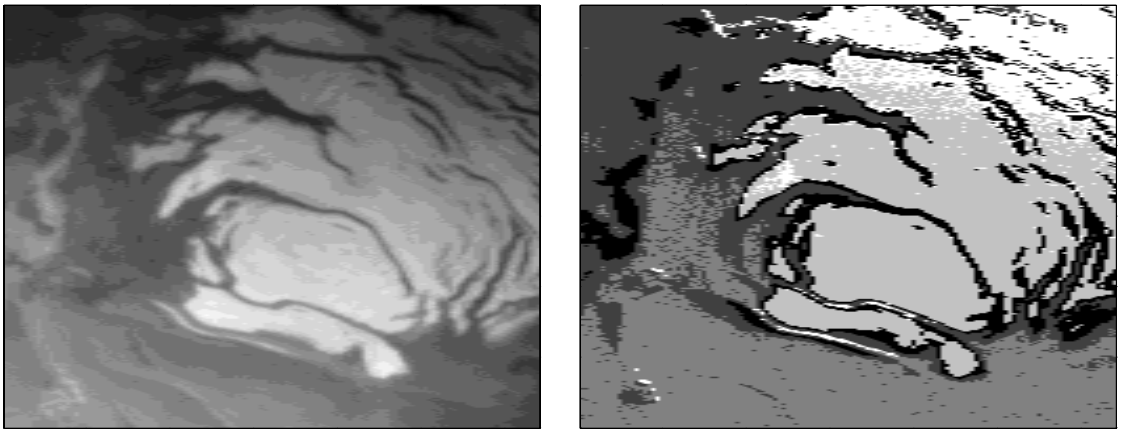

FIG. 7 – Catégorisation de la composition de la surface de Mars grâce à l'HDDC : à gauche, image de la zone étudiée et, à droite, segmentation par HDDC des données hyperspectrales de dimension 256 associées à l'image.

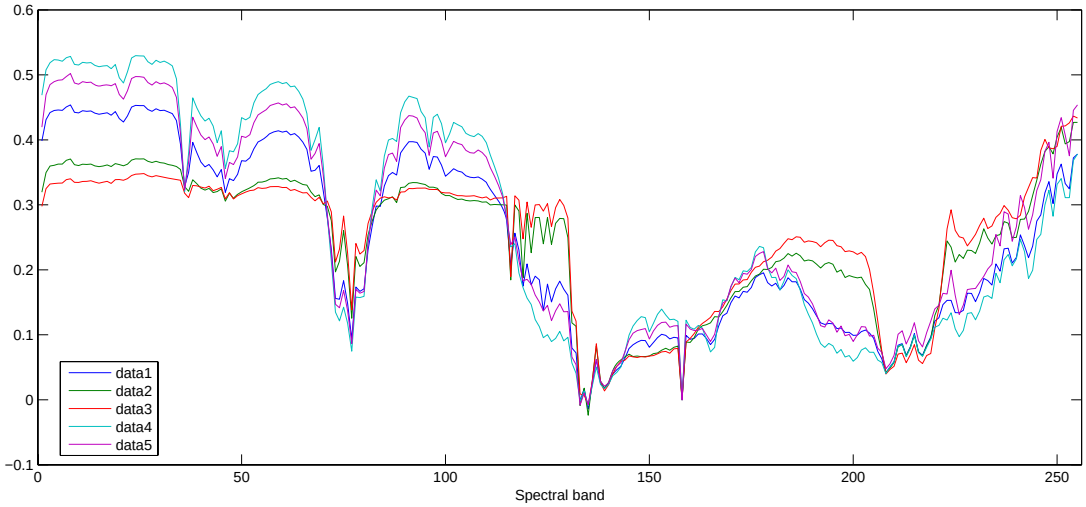

FIG. 8 – Moyennes spectrales des 5 classes minéralogiques trouvées grâce à l'HDDC.

marque que la classification sur composantes principales n'améliore pas les résultats des méthodes classiques. En revanche, on peut noter que la réduction de dimension permet d'améliorer significativement le résultat de classification du modèle full-GMM. La méthode de sélection de variables VS-GMM obtient un taux de classification correcte égal à 0.925 sur variables originales et à 0.935 sur composantes principales et devance ainsi les méthodes classiques. Enfin, l'HDDC domine cette étude en obtenant un taux de classification correcte égal à 0.95 sur variables originales et sur composantes principales. On peut déduire de la comparaison des résultats de l'HDDC avec ceux de VS-GMM que le fait de ne pas supprimer de variables et de modéliser chaque groupe dans son espace propre est la meilleure approche pour la classification. Notons au passage que l'HDDC est également capable de surclasser les méthodes existantes sur des jeux de données dont la dimension est plutôt faible.

5.3 Application à la caractérisation de la surface de Mars

Nous nous intéressons à présent au problème de la catégorisation d'images hyper-spectrales du sol de la planète Mars pour lequel les données sont à la fois de grande dimension et en très grand nombre. L'imagerie hyper-spectrale visible et infrarouge est une technique de télé-détection clef pour l'étude et le suivi des planètes du système solaire. Les spectromètres imageurs intégrés dans un nombre croissant de satellites génèrent des images hyper-spectrales à trois composantes (deux composantes spatiales et une spectrale). Les données, mises à notre disposition par le laboratoire de Planétologie de Grenoble, ont été acquises par l'imageur OMEGA. Cet imageur a observé le sol de la planète Mars avec une résolution spatiale variant entre 300 et 3000 mètres en fonction de l'altitude du satellite. Il a acquis pour chaque pixel observé les spectres dont les longueurs d'ondes vont de 0.36 à 5.2 μm et stocké ces informations dans un vecteur de 256 dimensions [5]. Le but de cette étude préliminaire est de caractériser la composition de la surface du sol martien en affectant chacun des pixels observés à une des 5 classes minéralogiques indiquées par les experts. Pour cette expérimentation, visant à vérifier l'aptitude de nos méthodes de classification à traiter de telles données, nous avons considéré une image de taille 300×128 pixels de la surface de la planète Mars dont chacun des 38 400 pixels est décrit par 256 variables. L'image de gauche de la Figure 7 représente la zone étudiée. L'image de droite de la Figure 7 montre la segmentation obtenue avec le modèle $[a_i b_i Q_i d_i]$ de l'HDDC. On peut tout d'abord observer que la segmentation fournie par l'HDDC est très satisfaisante sur une grande partie de l'image. Les résultats insuffisants de la partie supérieure droite de l'image sont dus à la courbure de la planète et peuvent être corrigés. Les experts du laboratoire de Planétologie de Grenoble ont particulièrement apprécié que notre méthode soit capable de détecter le mélange de glace et de carbonate (liseré noir) présent autour des zone de glaces (zones claires de l'image). La Figure 8 présente les moyennes spectrales des 5 classes. A partir de cette information, les experts peuvent déterminer avec précision la composition minéralogique de chacune des classes. Cette étude a démontré que notre méthode de *clustering* HDDC est capable de traiter efficacement des bases de données réelles de grande dimension et de grande taille. De plus, cette étude préliminaire a été réalisé sans prendre en compte les relations spatiales existantes entre les pixels et gageons que la prise en compte de ces relations améliore encore la segmentation. Nous envisageons de prendre en compte ces relations spatiales en utilisant l'approche qui combine l'HDDC à la modélisation par champs de Markov cachés et qui a donné des résultats prometteurs en reconnaissance de textures [7].

6 Conclusion

Nous avons, dans cet article, présenté une famille de modèles gaussiens pour les données de grande dimension. Ces modèles font l'hypothèse que les données de grande dimension vivent dans des sous-espaces de dimensions intrinsèques inférieures à la dimension de l'espace original et que les données de classes différentes vivent dans des sous-espaces différents dont les dimensions intrinsèques peuvent être aussi différentes. En forçant certains paramètres à être communs dans une même classe ou entre les classes, une famille de 28 modèles gaussiens adaptés aux données de grande dimension a été présentée, allant du modèle le plus général au modèle le plus parcimonieux. Ces modèles gaussiens ont été ensuite utilisés pour la discrimination et la classification automatique de données

de grande dimension. Les classifieurs associés à ces modèles sont baptisés respectivement *High Dimensional Discriminant Analysis* (HDDA) et *High Dimensional Data Clustering* (HDDC) et leur construction se base sur l'estimation des paramètres du modèle par la méthode du maximum de vraisemblance ou par l'algorithme EM. La nature des modèles présentés permet aux méthodes HDDA et HDDC de ne pas être perturbées par le mauvais conditionnement ou la singularité des matrices de covariance empiriques des classes et d'être efficaces en terme de temps de calcul. Des expériences numériques sur données simulées et réelles ont montré que les méthodes HDDA et HDDC sont aussi performantes en grande dimension qu'en dimension faible. Huit des modèles présentés dans cet article sont implantés dans le logiciel de classification MixMod et peuvent donc être mis en œuvre aisément sur tous types de données quantitatives.

Remerciements

Les auteurs souhaitent remercier Cordelia Schmid (INRIA Rhône-Alpes) pour ses conseils opportuns sur ce travail et Sylvain Douté (Lab. de Planétologie de Grenoble) pour avoir mis à notre disposition les données de catégorisation de la surface martienne ainsi que son expertise sur le sujet.

Références

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high-dimensional data for data mining application. In *ACM SIGMOD International Conference on Management of Data*, pages 94–105, 1998.

[2] J. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49 :803–821, 1993.

[3] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[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. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, and S. Girard. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research*, to appear, 2009.

[6] J.C. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure. I. Linear structure : fuzzy c-lines, II : Fuzzy c-varieties and convex combinations thereof. *SIAM J. Applied Mathematics*, 40 :339–357 and 38–372, 1981.

[7] J. Blanchet and C. Bouveyron. Modèle markovien caché pour la classification supervisée de données de grande dimension spatialement corrélées. In *38èmes Journées de Statistique de la Société Française de Statistique*, Clamart, France, 2006.

[8] L. Bocci, D. Vicari, and M. Vichi. A mixture model for the classification of three-way proximity data. *Computational Statistics and Data Analysis*, 50(7) :1625–1654, 2006.

[9] H.-H. Bock. The equivalence of two extremal problems and its application to the iterative classification of multivariate data. In *Mathematisches Forschungsinstitut*, 1969.

[10] H.-H. Bock. *Automatische Klassifikation*. Vandenhoeck and Ruprecht, Göttingen, 1974.

[11] H.-H. Bock. On the interface between cluster analysis, principal component clustering, and multidimensional scaling. In H.Bozdogan and A.K. Gupta, editors, *Multivariate statistical modeling and data analysis*, pages 7–34. Reidel, Dordrecht, 1987.

[12] H.-H. Bock. Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, 23(1) :5–28, 1996.

[13] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52 :502–519, 2007.

[14] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Discriminant Analysis. *Communications in Statistics : Theory and Methods*, 36(14) :2607–2623, 2007.

[15] R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2) :245–276, 1966.

[16] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14 :315–332, 1992.

[17] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *International Journal of Pattern Recognition*, 28(5) :781–793, 1995.

[18] G. De Soete and J.D. Carroll. K-means clustering in low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, *New approaches in classification and data analysis*, pages 212–219. Springer-Verlag, Heidelberg, 1994.

[19] P. Demartines and J. Hérault. Curvilinear Component Analysis : a self-organizing neural network for non linear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1) :148–154, 1997.

[20] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[21] W.S. DeSarbo and W.L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5 :249–282, 1988.

[22] E. Diday. Introduction à l'analyse factorielle typologique. *Revue de statistique appliquée*, 22 :29–38, 1974.

[23] 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 :169–184, 1986.

[24] L. Flury, B. Boukai, and B. Flury. The discrimination subspace model. *Journal of American Statistical Association*, 92(438) :758–766, 1997.

[25] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, 97 :611–631, 2002.

[26] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2) :145–167, 2000.

[27] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[28] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 :502–516, 1989.

[29] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[30] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, New York, 1995.

[31] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[32] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Interscience, New York, 1997.

[33] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.

[34] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41 :379–388, 2003.

[35] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data : a review. *SIGKDD Explor. Newsl.*, 6(1) :90–105, 2004.

[36] T. Pavlenko. On feature selection, curse of dimensionality and error probability in discriminant analysis. *Journal of Statistical Planning and Inference*, 115 :565–584, 2003.

[37] T. Pavlenko and D. Von Rosen. Effect of dimensionality on discrimination. *Statistics*, 35(3) :191–213, 2001.

[38] R.E. Quandt and J.B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73 :730–752, 1978.

[39] A. Raftery and N. Dean. Variable Selection for Model-Based Clustering. *Journal of American Statistical Association*, 101(473) :168–178, 2006.

[40] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500) :2323–2326, 2000.

[41] G. Saporta. *Probabilités, analyses des données et statistiques (2ème édition)*. Editions Technip, 2006.

[42] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 :1299–1319, 1998.

[43] J. Schott. Dimensionality reduction in quadratic discriminant analysis . *Computational Statistics and Data Analysis*, 66 :161–174, 1993.

[44] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464, 1978.

[45] D. Scott and J. Thompson. Probability density estimation in higher dimensions. In *Fifteenth Symposium in the Interface*, pages 173–179, 1983.

[46] J. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for non linear dimensionality reduction. *Science*, 290(5500) :2319–2323, 2000.

[47] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2) :443–482, 1999.