DISCRIMINATION DE COURBES PAR REGRESSION INVERSE FONCTIONNELLE

LOUIS FERRÉ ET NATHALIE VILLA

RÉSUMÉ. Les méthodes de régression inverse telles que la SIR (Li,1991) ont été développées dans le domaine de la régression multivariée pour éviter le célèbre fléau de la dimension. Elles ont été récemment étendues aux données fonctionnelles. Plusieurs approches ont été proposées et nous présentons ici un article de synthèse et de comparaison en abordant le cas où la variable réponse est un vecteur d'indicatrice d'appartenance à des classes. Nous montrons qu'alors la régression inverse conduit à une méthode de discrimination dont la pertinence est établie sur des données réelles et simulées.

Mots clefs : Discrimination, Données Fonctionnelles, Regression Inverse, Régression non paramétrique.

1. INTRODUCTION

L'analyse discriminante est une méthode éprouvée qui a été largement étudiée et étendue à différents contextes depuis sa découverte par Fisher. Les domaines d'application variés dans lesquels les problèmes de discrimination se rencontrent expliquent sans doute son succès. Elle se définit comme une méthode de classification supervisée qui consiste à classer des individus sur la base de variables explicatives et d'un échantillon d'apprentissage pour lequel à la fois ces variables et l'affectation aux classes sont connues.

Notons X la variable explicative, J le nombre de classes, C la variable identifiant les classes et $\mu_j = E(X|C = j)$ pour $j = 1, ..., J$. Le problème essentiel est celui de l'affectation des individus aux classes. Schématiquement, l'affectation peut s'opérer soit en utilisant des arguments géométriques, soit des arguments probabilistes. Dans le premier cas, affecter x à la classe j signifie que la distance de x au centre de gravité de la classe j est minimale, i.e., $d^2(x, \mu_j)$ est minimale en j pour une certaine distance d, habituellement la métrique de Mahalanobis. Cette règle peut s'appliquer directement aux données ou bien après réduction de la dimension par Analyse Factorielle Discriminante. Dans le deuxième cas, il s'agit de maximiser la probabilité $P(C = j|x)$ parmi toutes les valeurs de j. Au niveau statistique, tout le travail consiste à estimer cette probabilité. D'un point de vue paramétrique, la formule de Bayes fournit une réponse pour des modèles gaussiens et il est bien connu que cette règle d'affection est alors une version pénalisée de la règle géométrique. Mais on peut également estimer cette probabilité conditionnelle de façon non-paramétrique, voir e.g. Hand (1982). Cela bien sûr lorsque le régresseur est multivarié. Mais que se passe-t-il si le régresseur est fonctionnel ?

Comme la plupart des méthodes multivariées (voir, e.g., Dauxois et Pousse (1976) ou plus récemment Ramsay and Silverman (1997)), l'analyse discriminante a été

étendue au cas fonctionnel, moyennant cependant quelques adaptations. En particulier, si on considère le problème de l'affectation probabiliste, James and Hastie (2001) considérent le problème sous des hypothèses de normalité alors que Ferraty et Vieu (2003) aborde le problème sous l'angle de la régression non paramétrique.

Nous proposons ici une méthode sans hypothèse de loi. Elle est basée sur l'estimation du vecteur des probabilités $P = (P(C = j|X))_{j=1,...,J}$ à partir d'un modèle semi-paramétrique. Si on note Y le vecteur aléatoire de R^j tel que $Y = (Y_1, ..., Y_J)$ avec $Y_j = I_{[C=j]}$ où I est la fonction indicatrice de $[C = j]$ et X une variable aléatoire fonctionnelle, on obtient très simplement que :

$$P = E(Y|X). \tag{1}$$

On pose alors le modèle suivant :

$$P = f(< \beta_1, X >, ..., < \beta_d, X >) \tag{2}$$

où f est une fonction de R^d dans R^J, les β sont d fonctions définies sur le même ensemble que X. En fait, X est un processus stochastique continu $X(t)$ défini sur un intervalle I de R. On supposera que les fonctions X sont de carré intégrable et on considérera le produit scalaire sur L_I^2, ce qui signifie que $< \beta_k, X >= \int_I X(t)\beta_k(t)dt$.

Ce modèle est un modèle de réduction de dimension ; c'est une façon d'écrire que Y dépend de X uniquement au travers de sa projection sur un sous-espace d dimensionnel de L_I^2, engendré par les d vecteurs linéairement indépendants, $\beta_1, ..., \beta_d$. C'est, en ce sens, un espace "exhaustif" qui porte le nom d'espace EDR (pour Effective Dimension Reduction) dans la littérature sur la régression inverse (Li, 1991) ou central dans Cook et Yin(2001). S'agissant de données fonctionnelles, cet espace EDR va notamment permettre d'exhiber une base "optimale" au sens de la régression sur laquelle seront projetées les données avant de procéder à une autre analyse. L'estimation de P va dépendre de la façon dont est estimé cet espace. Dans la Section 2, nous verrons que si X admet un opérateur de covariance, la solution dérive directement des résultats de Dauxois et al. (2001) et que l'estimation de l'espace repose alors sur la décomposition spectrale de l'opérateur $\Gamma_X^{-1}\Gamma_{E(X|Y)}$, où Γ_Z désigne l'opérateur de covariance d'une variable fonctionnelle Z.

Pour un échantillon i.i.d. de taille n du couple (Y, X), $(Y_i, X_i)_{i=1,...,n}$, on déduit aisément des estimateurs convergents de Γ_X and $\Gamma_{E(X|Y)}$. Malheureusement, le premier opérateur n'est pas borné de sorte que son estimateur empirique est mal conditionné. Cette situation est bien connue en statistique fonctionnelle et plusieurs solutions ont été proposées pour contourner ce problème. Pour l'essentiel, il s'agit de méthodes de *régularisation* ou de *filtrage*. Dans le premier cas, on s'applique à charger la diagonale de l'estimateur de l'opérateur de covariance. Par exemple, la Ridge-regression (Wold,1975) est une méthode ancienne qui a été appliqués à l'analyse discriminate par Di Pillo (1979) et Friedman (1989). Une alternative consiste à utiliser un critère pénalisé comme dans l'analyse discriminante "fléxible", Hastie et al. (1994), ou dans l'analyse discriminante pénalisée, Hastie et al. (1995). Notons au passage que ces deux approches reposent sur la méthode d'" optimal scoring". On peut citer également les travaux de Leurgans, Moyeed and Silverman (1993) sur l'analyse des canoniques pénalisée en raison des liens étroits entre l'analyse canonique et l'analyse discriminante.

Dans James and Hastie (2001), une méthode de *filtrage* est utilisée. Elle consiste à projeter les données sur une base préalablement choisie, par exemple, une base d'ondelette, de polynômes orthogonaux, de polynômes trigonométriques ou de splines. L'avantage de cette approche est qu'elle autorise le traitement d'observations faites à des instants différents. L'inconvénient est que le choix des éléments de la base n'est pas toujours aisé.

Concernant la régression inverse fonctionnelle, plusieurs solutions ont été envisagées. Ainsi, Ferré and Yao (2003) utilisent une méthode de filtrage en projetant les données sur une base formée des premières fonctions propres de l'opérateur de covariance de X. Elle reprend l'idée développée par Bosq (1991) pour des modèles AR1. Ferré et Villa (2004) utilisent une méthode de régularisation. Enfin, pour éviter l'inversion, Ferré et Yao (2004) déduisent l'espace EDR des vecteurs propres de l'opérateur $\Gamma^{+}_{E(X|Y)}\Gamma_X$. Après avoir rappellé en Section 2 ce qu'est la régression inverse fonctionnelle, nous présenterons succinctement en Section 3 ces différentes méthodes en montrant comment elles s'inscrivent dans le cadre de l'analyse discriminante.

La Section 4 est elle consacrée à la régle d'affectation utilisée et nous terminons par la Section 5 où les méthodes seront mises en oeuvre et comparées sur des données réelles ou simulées.

2. La regression inverse

Soit X une variable aléatoire à valeur dans l'espace des fonctions de carré intégrable L^2_I, où I est un intervalle de R et Y une variable aléatoire à valeur dans R^J.

La régression inverse fonctionnelle s'appuie sur le modèle suivant :

$$Y = f(< \beta_1, X >, ..., < \beta_d, X >) + \epsilon \ , \tag{3}$$

où $\beta_1, ..., \beta_d$ sont des vecteurs de L^2_I linéairement indépendants, f est une fonction, inconnue, de R^d dans R^J et ϵ est une variable aléatoire dans R^J non-corrélée avec X. D'un côté, ce modèle (3) est un cas particulier du modèle semi-parametrique pour variables hilbertiennes présenté dans Dauxois et al. (2001) dont nous allons exploiter les propriétés. D'un autre côté, il admet comme cas particulier la situation où Y est un vecteur aléatoire de R^j tel que $Y = (Y^{(1)}, ..., Y^{(J)})$ avec $Y^{(j)} = I_{[C=j]}$, où I est la fonction indicatrice de $[C = j]$, et il sera donc un modèle pertinent pour l'analyse discriminante.

Si notre approche ne repose sur aucune hypothèse de loi, il est cependant nécessaires de supposer que :

Hypothèse H-1 pour tout $b \in L^2_I$, si on pose $B' = (< \beta_1, X >, ..., < \beta_K, X >)$, alors $E(< b, X > |B)$ est linéaire en B ;

Hypothèse H-2 $E(\| X \|^4) < \infty)$.

L'hypothèse *H-1* est à la fois un cas particulier de l'hypothèse *H-1* de Dauxois et al. (2001) et la version fonctionnelle de l'hypothèse 1.6 of Li (1991). Notons qu'elle est vérifiée notamment si X est une variable fonctionnelle gaussienne ou plus généralement elliptique.

L'hypothèse H-2 assure l'existence de l'espérance de X, $E(X)$, notée μ par la suite et de son opérateur de covariance noté Γ_X. Cela permet également de définir $\mu_j = E(X|C = j)$ pour tout $j = 1, ..., J$ et $\Gamma_{E(X|Y)}$, l'opérateur de covariance de $E(X|Y)$. Par ailleurs, on supposera tout au long de l'article que :

Hypothèse H-3 Γ_X est définie positive.

Soient η_k les vecteurs propres de l'opérateur $\Gamma_X^{1/2}\Gamma_{E(X|Y)}\Gamma_X^{1/2}$, on pose $b_k = \Gamma_X^{-1/2}\eta_k$. Pour garantir, l'existence des vecteurs b_k ainsi définis, nous supposerons que (voir Ferré et Yao, 2004) :

Hypothèse H-4 Si $X = \sum_{i=1}^{\infty} \xi_i u_i$ est la décomposition de Karunen-Loeve de X, alors $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{E(\xi_i|Y)E(\xi_i|Y)}{\delta_i^2\delta_j} < \infty$ où $\delta_i = E(\xi_i)$.

En utilisant les résultats de Dauxois et al.(2001), on vérifie aisément que :

Théorème 2.1. *Sous les hypothèses H-1, H-2, H-3, H-4, $sp\{b_1,...,b_d\}$ est inclus dans l'espace EDR.*

On en déduit alors que l'estimation de l'espace EDR s'obtient à partir de celle de $sp\{b_1,...,b_d\}$. L'analogie avec la SIR et en particulier avec la SIR fonctionnelle est évidente mais deux points méritent d'être précisés. Tout d'abord, l'estimation de $\Gamma_{E(X|Y)}$ ne nécessite pas de tranchage et s'obtient directement par la matrice de covariance inter-groupe. Ensuite, la variable Y est ici multivariée alors qu'en SIR ou FSIR, elle est généralement univariée (voir cependant, pour la SIR Hsing(1999) et Li et al.(2003)).

Si nous considérons ici le cas fonctionnel, notre approche s'applique également au cas multivarié. L'utilisation des méthodes de réduction de dimension a été récemment considérée en analyse discriminante multivariée. En effet, Cook et Yin(2001) considèrent un modèle comparable au modèle (3), mais dans lequel P est remplacé par la coordonnée maximale de P, revenant ainsi à un modèle univarié. Ils utilisent ensuite la méthode SIR ou SAVE (Cook, 1991) pour estimer ce sous-espace. De même, Hernandez et Velilla (2001) estiment l'espace central qui maximise la règle de Bayes. Leur technique repose sur la maximisation d'un critère basé sur l'entropie et conduit à une procédure beaucoup plus lourde que la méthode présentée ici dans le cas multivarié et elle n'a pas, à ce jour, d'équivalent dans le cas fonctionnel.

Les estimateurs des vecteurs de base de la méthode par réduction de dimension sont identiques aux fonctions discriminantes de l'analyse discriminante linéaire. Même si, *a priori,* ces deux problèmes ne sont pas à la base semblables, la relation entre l'analyse discriminante linéaire et la régression inverse provient du fait que chacune d'elle se ramène, pour la première direction, au problème de maximisation du critère de Rayleigh :

$$max \frac{< \Gamma_{E(X|Y)} b, b >}{< \Gamma_X b, b >}, \tag{4}$$

les autres directions étant solutions de problèmes identiques sous contraintes d'orthogonalité. Ceci est vrai dans le cas fonctionnel ou multivarié.

3. Estimation des paramètres

A partir d'un échantillon i.i.d. de taille n, (X_i, Y_i), pour $i = 1, ..., n$, les estimateurs de l'espace central se déduisent des estimateurs suivants. On estime $\Gamma_{E(X|Y)}$ par la matrice de covariance inter groupes,

$$\Gamma^n_{E(X|Y)} = \sum_{j=1}^{J} \frac{n_i}{n}(\hat{\mu}_j - \bar{X}) \otimes (\hat{\mu}_j - \bar{X}),$$

où $n_j = \sum_{i=1}^n Y_i^{(j)}$ et $\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^n X_i Y_i^{(j)}$. L'opérateur Γ_X est estimé par l'opérateur empirique, $\Gamma_X^n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}) \otimes (X_i - \bar{X}))$, où $\bar{X}$ est la moyenne empirique de X. Cependant, sous l'hypothèse H-2, Γ_X est un opérateur de Hilbert-Schmidt est n'est donc pas inversible dans L_I^2. De plus, si l'on retreint Γ_X à son image, l'opérateur ainsi obtenu est inversible, mais son inverse n'est pas borné. Ainsi, la matrice Γ_X^n sera mal-conditionnée et il convient d'utiliser des stratégies pour contourner ce problème. Nous présentons ci-dessous plusieurs solutions proposéees dans le cadre de la régression inverse.

3.1. **Une solution de filtrage.** Nous présentons ici l'approche de Ferré et Yao (2003) : la "Fonctional Sliced Inverse regression", FSIR. Elle consiste à projeter les données dans une base des vecteurs propres de l'opérateur Γ_X. Cela conduit à un choix "objectif" de la base de projection alors que l'utilisation de bases orthogonales comme celle de Fourier, de fonctions splines, d'ondelettes, ne permettent pas des choix toujours pertinents des éléments de la base. Soit $(k_n)_{n\in N}$ une suite non convergente d'entiers. Pour tout n, on note Π_{k_n} le projecteur propre associé aux k_n plus grande valeurs propres de Γ_X^n. L'estimation de l'espace EDR s'obtient par décomposition spectrale de l'opérateur $((\Pi_{k_n}\Gamma_X^n\Pi_{k_n})^+)^{1/2}\Gamma_{E(X|Y)}^n((\Pi_{k_n}\Gamma_X^n\Pi_{k/_n})^+)^{1/2}$ où la notation "+" est utilisée pour représenter l'inverse généralisé d'une matrice. La consistance de l'estimateur de l'espace EDR a été étudiée dans Ferré and Yao (2003) lorsque Y est une variable aléatoire réelle. La convergence de $\Gamma_{E(X|Y)}^n$ vers $\Gamma_{E(X|Y)}$ permet d'étendre directement ce résultat au cas où Y est un vecteur d'indicatrice. Ces résultats reposent sur des hypothèses sur les valeurs propres de Γ_X. Grossièrement, celles-ci ne doivent pas tendre vers 0 trop rapidement. Précisons que ces hypothèses sont vérifiées, en particulier, si X est un mouvement brownien. La mise en oeuvre de cette méthode nécessite de sélectionner une valeur convenable pour k_n.

3.2. **Une solution basée sur un inverse généralisé de $\Gamma_{E(X|Y)}^n$.** Cette solution repose sur le fait que sous le modèle (3), $\Gamma_{E(X|Y)}$ est un opérateur de rang fini. Ainsi, $(\Gamma_X)^{-1/2}\Gamma_{E(X|Y)}(\Gamma_X)^{-1/2}$ est lui-même de rang fini et l'espace propre engendré par ses valeurs propres non-nulles est identique à celui celui de son inverse généralisé. Le problème qui se pose ici est que cela nécessite la connaissance *a priori* de la dimension d de l'espace EDR. Si on note $(\Gamma_{E(X|Y)}^n)^{+d}$ l'inverse généralise de $\Gamma_{E(X|Y)}^n$ tronqué aux d plus grandes valeurs propres, l'espace EDR est estimé à partir de la diagonalisation de la matrice $((\Gamma_X^n)^{1/2}(\Gamma_{E(X|Y)}^n)^{+d}(\Gamma_X^n)^{1/2})$.

Cette méthode a été proposée par Ferré et Yao (2004) dans le cadre de la régression inverse lorsque Y est une variable aléatoire réelle et les auteurs établissent la consistance de l'estimateur de l'espace EDR sous des hypothèses assez faibles. Ces résultats s'entendent là encore à notre contexte. L' avantage principal de cette approche est qu'elle est particulièrement simple à mettre en oeuvre. Cependant, elle nécessite l'estimation de d pour calculer $(\Gamma_{E(X|Y)}^n)^{+d}$. Si dans Ferré et Yao (2004), un critère lié à l'estimation de $(\Gamma_{E(X|Y)}^n)$ est proposé, nous estimons ici d à partir d'un critère prenant en compte l'ensemble de la procédure. Nous développerons ce point dans les applications.

3.3. **Une approche par régularistion.** Les méthodes de régularisation sont très communes pour le traitement des données fonctionnelles. Elles sont présentées

comme plus efficaces que les méthodes de filtrage. L'idée principale est de pénaliser l'opérateur de covariance en introduisant des contraintes de régularité sur les fonctions estimées par modification du critère optimisé lors de l'estimation.

Ici, rappelons-le, la première direction, b_1, est la solution du problème $max \frac{<\Gamma_{E(X|Y)}b,b>}{<\Gamma_X b,b>}$, la deuxième, b_2, celle de $max \frac{<\Gamma_{E(X|Y)}b,b>}{<\Gamma_X b,b>}$, sous la contrainte $< \Gamma_X b_1, b >= 0$, etc...

Pour prendre en compte les contraintes de régularité, nous introduisons un paramètre de lissage λ et nous considérons la procédure suivante :

- déterminer $\hat{b}_1$ tel que $\hat{b}_1$ maximise le critère $\frac{<\Gamma^n_{E(X|Y)}b,b>}{<\Gamma^n_X b,b>+\lambda<D^2 b,D^2 b>}$;
- puis déterminer $\hat{b}_2$ tel que $\hat{b}_2$ maximise $\frac{<\Gamma^n_{E(X|Y)}b,b>}{<\Gamma^n_X b,b>+\lambda<D^2 b,D^2 b>}$, sous la contrainte $< \Gamma^n_X b, \hat{b}_1 > +\lambda < D^2 b, D^2 \hat{b}_1 >= 0$;
- les autres vecteurs s'obtiennent par optimisation du critère sous des contraintes semblables.

La solution à ce problème est donnée par $(\hat{b}_1, ..., \hat{b}_d)$, vecteurs propres associés aux d plus grandes valeurs propres de la matrice $(\Gamma^n_X + \lambda D^4)^{-1}\Gamma^n_{E(X|Y)}$ et vérifiant $< (\Gamma^n_X + \lambda D^4)\hat{b}_i, \hat{b}_j >= \delta_{ij}$. La matrice D^4 est un estimateur de l'opérateur de différentiation d'ordre 4 et est calculée à partir d'un base de fonctions splines.

La consistance de ces estimateurs a été démontrée dans Ferré et Villa (2004) pour une variable réelle, mais elle reste également valable dans le cadre étudié ici.

4. Régle de classification

Dans le modèle (2), f est une fonction de lien définie de R^d dans R^J. Elle s' écrit $f = (f_1, ..., f_J)$ où, pour chaque u dans R^d et chaque j, $f_j(u) = P(C = j|U = u)$, avec U est la projection de X sur l'espace EDR et $f(x) = E(Y|X = x)$. Une fois l'espace EDR estimé, le problème de l'estimation de f se ramène à un problème de régression multivariée et l'estimation des probabilités d'appartenance aux groupes sachant X va s'obtenir par régression non paramétrique.

Toute méthode non paramétrique peut être, bien sûr, utilisée, mais nous emploierons ici des estimateurs à noyaux. On considère l'estimateur de Nadaraya-Watson :

$$\hat{f}_j(u) = \frac{\Sigma^n_{i=1} Y_i^{(j)} \Pi^d_{k=1} K(\frac{<b_k, X_i> - u}{h})}{\Sigma^n_{i=1} \Pi^d_{k=1} K(\frac{<b_k, X_i> - u}{h})} \tag{5}$$

où K est un noyau vérifiant $\int K(v)dv = 1$ et h est la fenêtre.

En fait, ce critère est une version non paramétrique de la règle de Bayes appliquée aux données projectées sur l'espace EDR. En effet, $\frac{\Sigma^n_{i=1} Y_i^{(j)} \Pi^d_{k=1} K(\frac{<b_k, X_i> - u}{h})}{nh}$ est l'estimateur à noyau de la densité de $(< b_1, X >, ..., < b_d, X >)$ conditionnelle à $Y^{(j)}$ et $\frac{\Sigma^n_{i=1} \Pi^d_{k=1} K(\frac{<b_k, X_i> - u}{h})}{nh}$ est l'estimateur de la densité conjointe de $(< b_1, X >, ..., < b_d, X >)$. Ainsi, la règle de classification conduit à maximiser en j un estimateur non paramétrique de $P(Y^{(j)}|(< b_1, X >, ..., < b_d, X >))$ dans l'esprit de ceux présentés dans Devroye et al. (1996).

L'utilisation d'estimateurs à noyau peut s'avérer, dans le cas de groupes trop nombreux, inefficace en raison du "fléau de la dimension". Si tel est le cas, il est possible de remplacer ces estimateurs à noyau par des réseaux de neurones, insensibles à ce problème, et dont la pertinence de leur association avec les méthodes de régression inverse est démontrée dans Ferré et Villa (2004). Cependant dans

la pratique, le nombre de groupes est généralement raisonnablement faible ce qui explique le choix effectué ici.

5. Applications

5.1. **Méthode.** Cette partie est consacrée à la mise en oeuvre des méthodes ci-dessus et à leur comparaison avec des méthodes concurrentes. Que les données soient réelles ou simulées, le mode opératoire présenté ci-après leur sera commun.

Pour obtenir les estimateurs en pratique, nous proposons de diviser l'échantillon en trois et de procéder de la façon suivante :

- tout d'abord estimer l'espace EDR sur le premier échantillon, dit échantillon d'apprentissage ;
- puis, déterminer la fenêtre et les paramètres des différents modèles par validation croisée sur un échantillon de contrôle ;
- enfin de déterminer le pourcentage de mal classés sur un échantillon test.

Cette façon de procéder présente les avantages suivants : tout d'abord, les deux premières étapes s'effectuant sur des échantillons indépendants, il est possible d'utiliser les résultats de convergence des estimateurs à noyau pour obtenir la convergence de l'estimateur de f. Ensuite, il est possible de considérer la dimension d comme un paramètre du modèle pour l'approche 2 et de réitérer la procédure avec différentes valeurs de d pour retenir celle qui minimise le taux de mal classés. Dans le cadre de notre étude, nous avons, tout d'abord et comme décrit ci-dessus, déterminé par validation croisée les paramètres de chaque modèle puis, nous considérons cinquante segmentations de l'échantillon en deux parties obtenues de façon aléatoire : le premier échantillon permet la mise en oeuvre des différentes méthodes comparées pour les paramètres optimaux et le second échantillon évalue la performance moyenne et la variabilité de chacune de ces méthodes. Notons que les paramètres optimaux ont été callés avant les répartitions aléatoires de l'échantillon pour éviter des explosions de temps de calcul liés au volume important des données. Nous désignerons par SIR-Np la méthode par projection, SIR-N celle par inverse généralisé et SIR-Nr celle par régularisation. Nous les comparerons avec RPDA (Ridge Penalized Discriminant Analysis) de Hastie et al. (1995) et avec la méthode à noyaux de Ferraty et Vieu (2003) désignée ici par NPCD-PCA.

5.2. **Données simulées : les "waveform data".** Nous considérons ici un jeu de données simulées qui est une sorte d'étalon pour la comparaison des méthodes de discrimination fonctionnelle. La base de donnnées est composée de 3000 courbes discrétisées en 21 points ($t = 1, 2, \ldots, 21$) qui sont issues de 3 familles différentes (1000 courbes par classes) :

(1) $t \to uh_1(t) + (1-u)h_2(t) + \epsilon(t)$;

(2) $t \to uh_1(t) + (1-u)h_3(t) + \epsilon(t)$;

(3) $t \to uh_2(t) + (1-u)h_3(t) + \epsilon(t)$

où u est une variable aléatoire de loi uniforme sur $[0;1]$, $\epsilon(t)$ est une variable aléatoire de loi normale centrée réduite et

$$h_1(t) = max(6 - |t-11|, 0),\ h_2(t) = h_1(t-4) \text{ et } h_3(t) = h_1(t+4)\ .$$

Des représentations de ces courbes sont données en Figure 1.

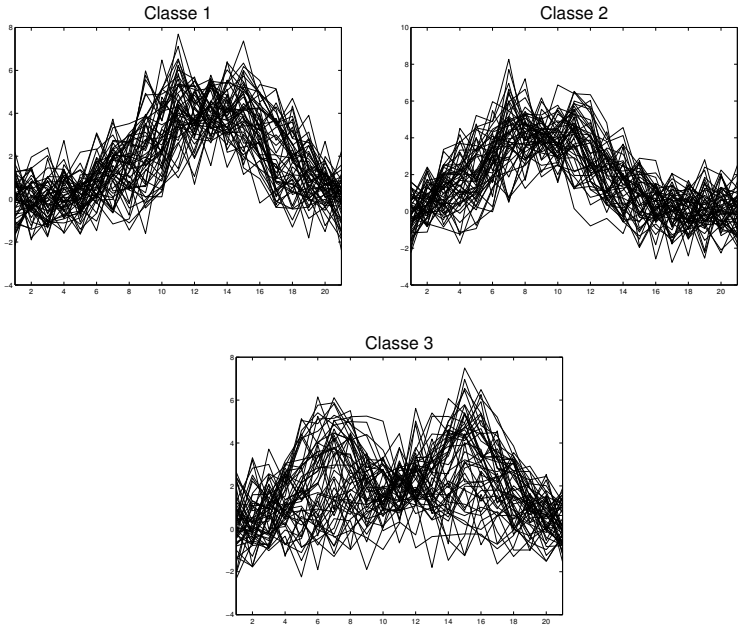

FIG. 1. Un échantillon de 50 courbes par classe

La base de données a été partagée en deux de manière aléatoire : un échantillon de 1500 individus (500 par classe) constituait la base d'apprentissage et un échantillon de 1500 individus (500 par classe) la base de test.

Les valeurs optimales pour chacunes des méthodes employées sont données dans le Tableau 1.

	Paramètre 1	Paramètre 2	Paramètre 3
SIR-Nr	$\lambda = 1$ (régularisation de Γ_X)	$d = 2$ (dimension SIR)	$h = 0{,}75$ (fenêtre du noyau)
SIR-Np	$k_n = 2$ (dimension ACP)	$d = 2$ (dimension SIR)	$h = 3$ (fenêtre du noyau)
SIR2-N	$d = 2$ (dimension SIR)	$h = 3$ (fenêtre du noyau)	
RPDA	$\lambda = 2$ (régularisation de Γ_X)	$d = 2$ (dimension AFD)	
NPCD-PCA	$k_N = 16$ (dimension ACP)	$h = 6$ (fenêtre du noyau)	

TAB. 1. Valeurs optimales

Les fonctions qui engendrent de l'espace EDR et l'espace dicriminant dans la RPDA sont données Figure 2. Ces fonctions propres sont très proches en ce qui concerne SIR2-N et SIR-Np et toutes deux sont assez voisines de celles de la RKDA.

On peut observer aussi que seule SIR-Nr conduit à des fonctions lisses ce qui est cohérent avec cette méthode.

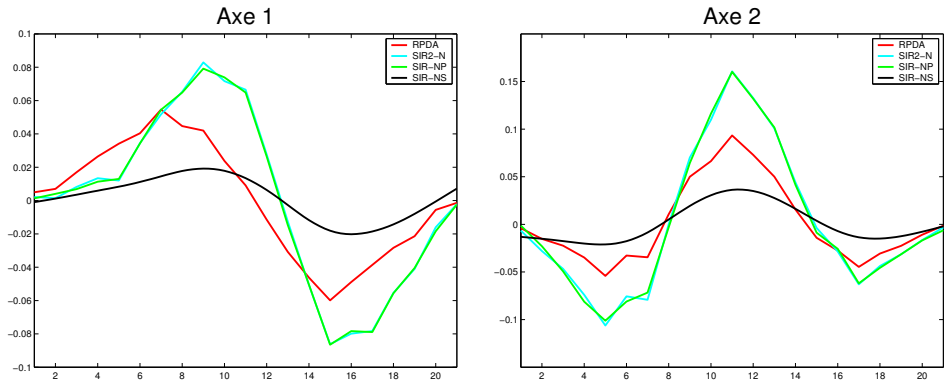

FIG. 2. Vecteurs de l'espace EDR

La projection du nuage de points sur ces espaces est donnée en Figure 3. Il est difficile de distinguer une différence entre les différents graphiques laissant ainsi présager de performances comparables dans les affectations. C'est en effet ce que l'on observe à la Figure 4 qui donne les boîtes à moustaches des taux d'erreurs de classements construites à partir des cinquantes simulations. On peut constater que la méthode RPDA est celle qui donne les plus mauvais résultats alors que les meilleurs sont obtenus par SIR2-N. La relative médiocrité des résultats de SIR-Nr s'explique sans doute par la difficulté d'interpoler convenablement les 21 points de discrétisation sur une base spline. Le Tableau 2 donne les caractéristiques du taux de mal classés. On peut y remarquer que les méthodes reposant sur la régression inverse fonctionnelle fournissent globalement des résultats qui les situent parmi les méthodes les plus performantes si on les compare avec ceux d'études similaires (Hernandez et Velilla (2001) indiquent que leur méthode RKDA a un taux d'erreur de 16,2 % et que le meilleur taux est obtenu pour un réseau de neurones avec 15,1 %).

	Moyennes	Médianes	Ecart type	1° quartile	Minimum
SIR-Nr	16,62 %	16,67 %	0,63 %	16,33 %	15,33 %
SIR-Np	16,15 %	16,33 %	0,77 %	15,73 %	14,20 %
SIR2-N	15,92 %	15,93 %	0,55 %	15,60 %	14,73 %
RPDA	18,38 %	18,40 %	0,68 %	18,00 %	17,07 %
NPCD-PCA	16,37 %	16,40 %	0,66 %	15,93 %	14,87 %

TAB. 2. Caractéristiques des taux de mal classés

5.3. **Reconnaissance de phonèmes.** La base de données est composée de 4509 log-périodogrammes (discrétisés en 256 points) qui correspondent aux enregistrements de 5 phonèmes différents dont des représentations sont données en Figure 5.

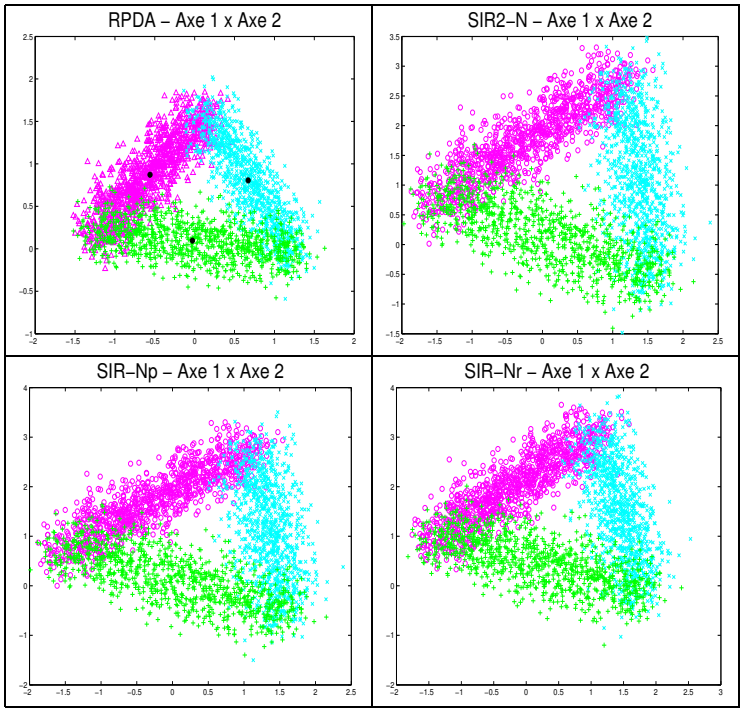

FIG. 3. Projection des données sur l'espace EDR

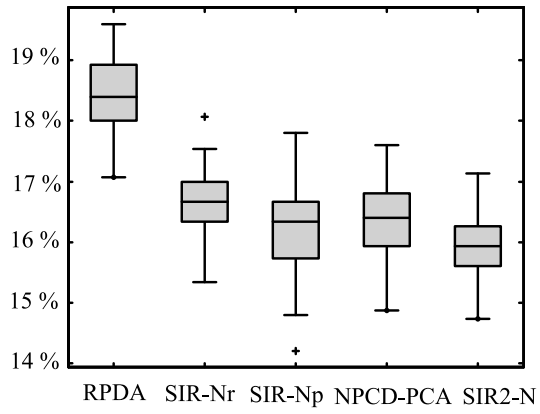

FIG. 4. Comparaison des taux d'erreur pour 50 échantillons

Les phonèmes enregistrés sont [sh] (872 log-périodogrammes), [iy] (1163 log-périodogrammes), [dcl] (757 log-périodogrammes), [aa] (695 log-périodogrammes) et [ao] (1022 log-périodogrammes). La base de données a été partagée en deux de manière aléatoire : un échantillon de 1735 individus (347 par classe) constituait la base d'apprentissage et un échantillon de 1735 individus (347 par classe) la base de test sur laquelle était calculée l'erreur correspondant à chaque méthode.

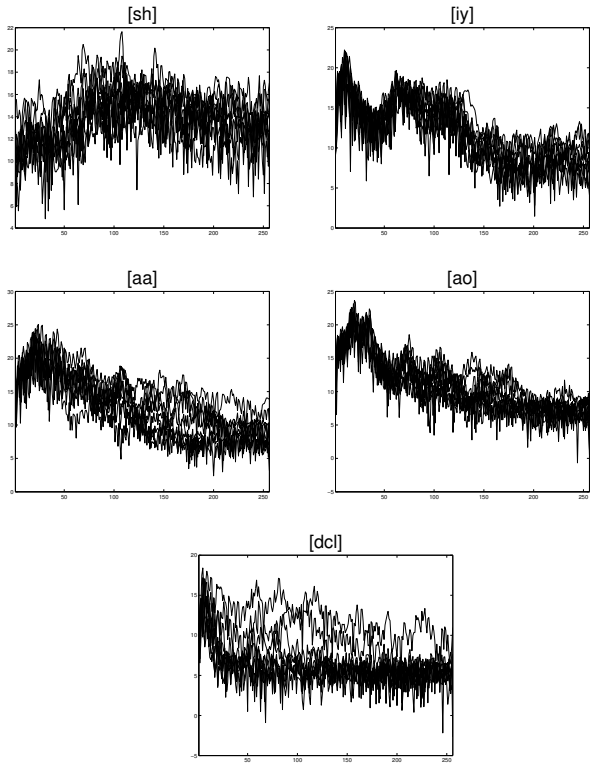

FIG. 5. Un échantillon de 10 log-périodogrammes par classe

Les vecteurs engendrant l'espace EDR sont représentés Figure 6. Les solutions les moins lisses sont fournies par la méthode SIR-Nr (en raison d'une faible valeur du paramètre optimal de régularisation) alors que les plus lisses sont obtenues par RPDA et SIR2-N. Les solutions fournies par les diverses méthodes testées sont très proches les unes des autres.

Les Figures 7, 8, 9 et 10, permettent de visualiser la projection des données sur l'espace EDR fournie, respectivement, par les méthodes SIR-Nr, SIR-Np, SIR2-N et RPDA. Elles font apparaître une bonne séparabilité linéaire des données projetées, laissant penser que le simple modèle RPDA fournira des taux d'erreur très satisfaisants. Concernant les méthodes SIR2-N et RPDA, les axes 1 et 2 permettent de séparer les phonèmes [sh] (en haut à droite), [ly] (au centre), [dcl] (en bas) et les phonèmes en "a", [aa] et [ao] qui eux sont confondus (à gauche). A l'inverse, SIR-Nr et SIR-Np, ne permettent pas de séparer [ly] et [dcl] sur l'axe 2 : en deux dimensions, SIR2-N et RPDA sont les plus discriminantes. Si l'axe 3 donne des résultats comparables d'une méthode à l'autre, en séparant bien [ly] et [dcl], l'axe 4 est le seul qui permet de discriminer les phonèmes [aa] et [ao]. La discrimination y est meilleure pour SIR-Nr et SIR-Np que pour RPDA et SIR-N, ce qui explique que, globalement, les premières ont des performances supérieures aux secondes comme l'indiquent le Tableau 3 et la Figure 11.

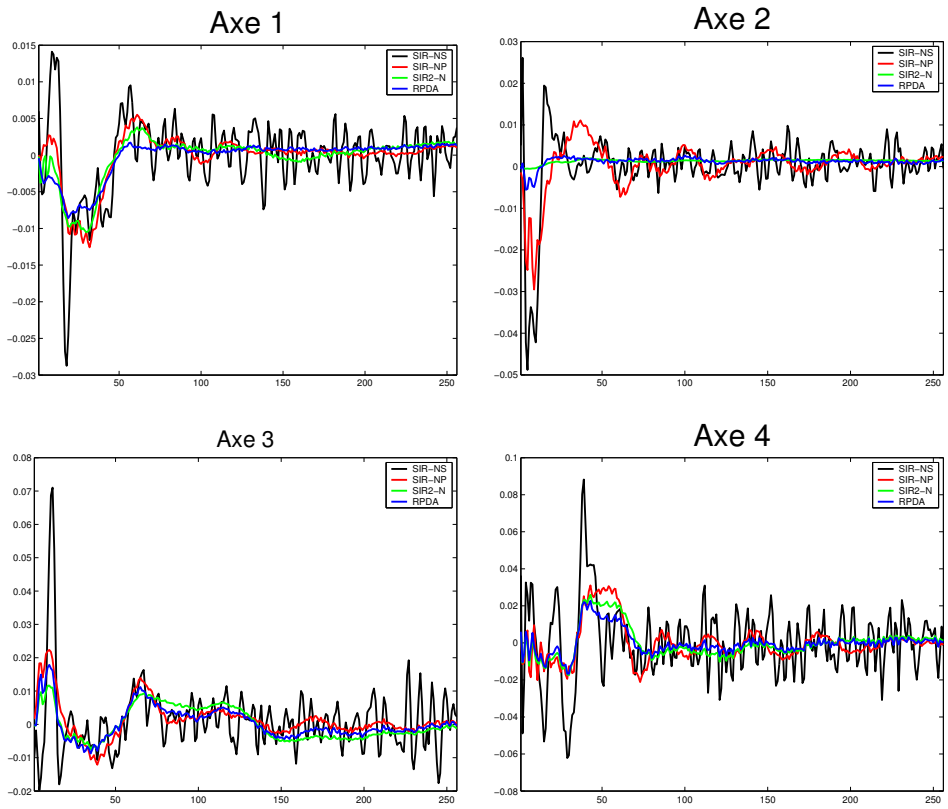

FIG. 6. Vecteurs de l'espace EDR pour les phonèmes.

	Moyennes	Médianes	Ecart type	1° quartile	Minimum
SIR-Nr	8,24 %	8,30 %	0,44 %	7,95 %	7,32 %
SIR-Np	8,45 %	8,44 %	0,51 %	8,01 %	7,32 %
SIR2-N	9,33 %	9,48 %	0,47 %	8,99 %	8,36 %
RPDA	9,04 %	9,05 %	0,52 %	8,65 %	7,84 %
NPCD-PCA	9,89 %	9,83 %	0,60 %	9,39 %	8,47 %

TAB. 3. Caractéristiques des taux de mal classés pour les phonèmes.

RÉFÉRENCES

[1] Bosq, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes. In : Roussas, G. (Ed.), Nonparametric Functional estimation and related Topics, NATO, ASI Series, pp. 509-529.

[2] Cook, R.D. (1991) Discussion of Li (1991) *J. Am. Statis. Ass.*, **86**,328-332.

[3] Cook, R.D. et Yin X.(2001) Dimension reduction and visualization in discriminant analysis, *Australian & New-Zealand Journal of Statistics,* **43**, 147-199.

[4] Dauxois, J. and Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. *Thèse Toulouse III.*

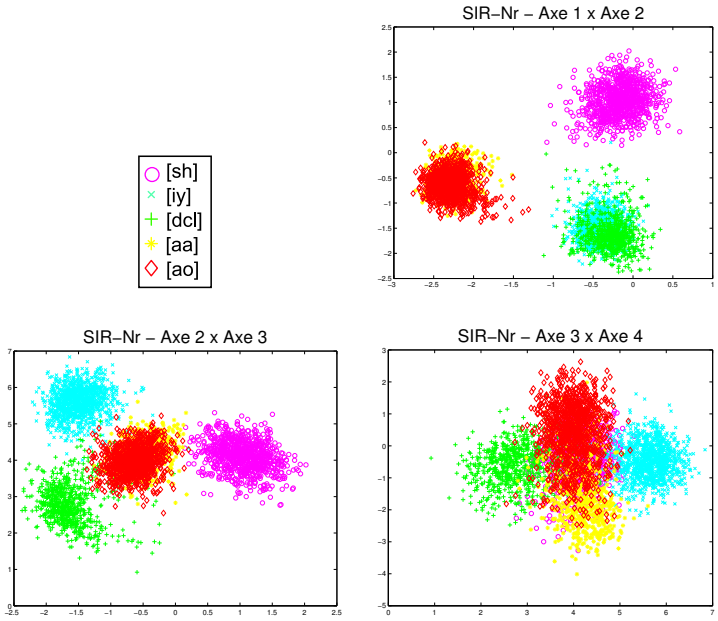

FIG. 7. Projection des données (SIR-Nr)

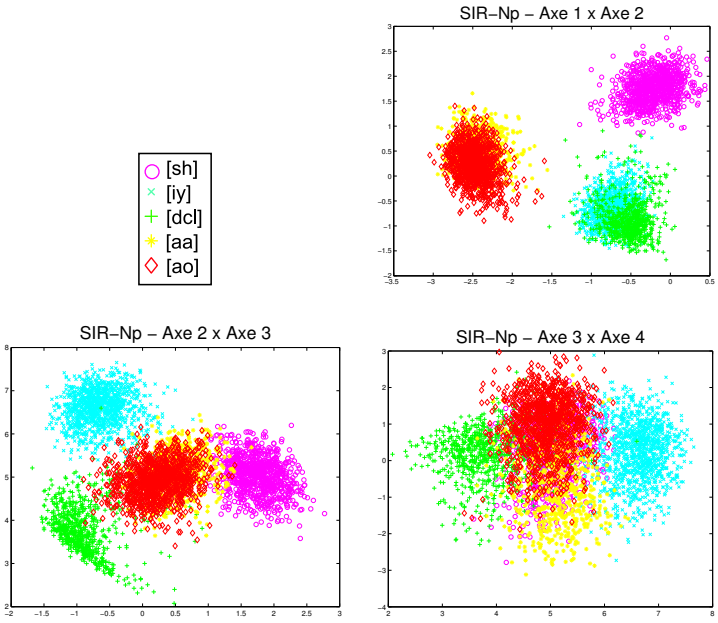

FIG. 8. Projection des données (SIR-Np)

[5] Dauxois, J., Ferré, L. and Yao, A.F. (2001) Un modèle semi-paramétrique pour variable aléatoire hilbertienne. *C.R. Acad. Sci. Paris*, **t.327**, série I, 947-952.

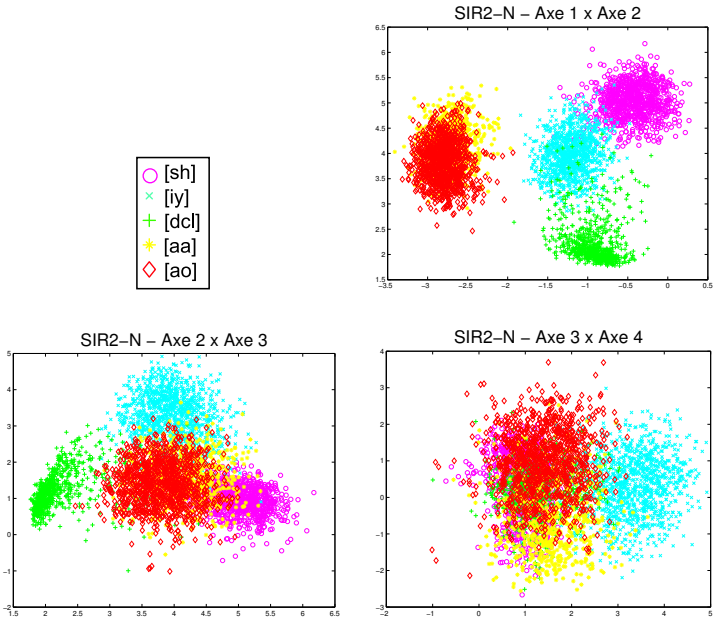

FIG. 9. Projection des données (SIR2-N)

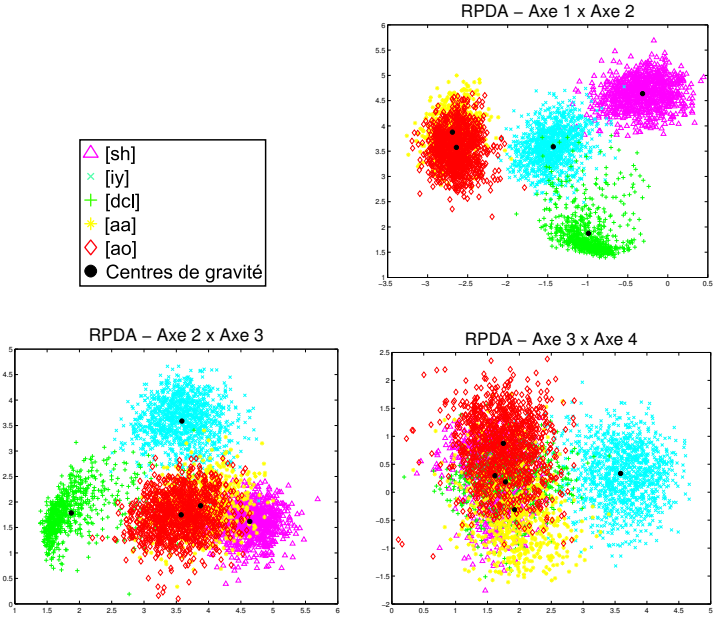

FIG. 10. Projection des données (RPDA)

[6] Devroye L., Györfi L. and Lugosi G. (1996) *A probabilistic theory for pattern recognition*, New-York : Springer-Verlag.

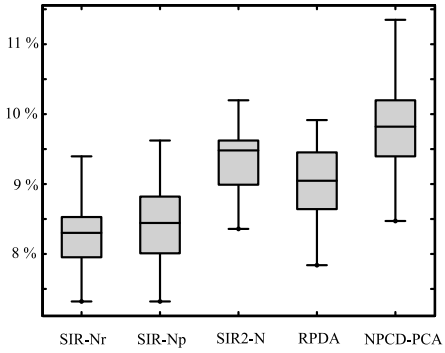

FIG. 11. Comparaison des taux d'erreur pour 50 échantillons

[7] DiPillo, P. (1979) Biased discriminant analysis : evalaution of the optimum probability of classification. *Comm. Statist. Theory Methods,* **8**, 1447-1458.
[8] Ferraty, F. and Vieu, P. (2003) Curves Discrimination : a Non Parametric Approach. Computational and Statistical Data Analysis, **44,** 161-173. *Computational Statistics*, **17**, 545-564.
[9] Ferré, L. and Villa, N. (2004) Multi-layer Neural Network with Functional Inputs. Preprint.
[10] Ferré, L. and Yao, A. F. (2003) Functional Sliced Inverse Regression analysis. *Statistics*, **37**, 475-488.
[11] Ferré, L. et Yao, A.F. (2004) Smoothed Functional Inverse Regression, Soumis à publication.
[12] Friedman, J. (1989) Regularized discriminant analysis. *J. Amer. Statist. Assoc.,* **84**, 165-175.
[13] Hand D.J. (1982) *Kernel discriminant analysis,* Research Studies Press/Wiley.
[14] Hastie T., Tibshirani R. and Buja A. (1994) Flexible Discriminant Analysis by optimal scoring, *J. Amer. Statist. Ass.,* **89,** 1255-1270.
[15] Hastie T., Buja A. and Tibshirani R. (1995) Penalized Discriminant Analysis, Annals of Statistics (23), p 73-102.
[16] Hernandez A. et Velilla S. (2001) Dimension reduction in nonparametric discriminant analysis, *Technical report.* **85**, 54-77.
[17] Hsing T.(1999) Nearest Neighbor Inverse Regression, Ann. Statist, 697-731.
[18] James G.M., Hastie T.J. and Sugar C.A. (2000). Principal Component models for sparse functional data. *Biometrika*, **87**, 3, 587-602.
[19] Leurgans , S.E., Moyeed, R.A. and Silverman, B.W. (1993). Canonical Correlation Analysis when the data are curves. *J.R.Statis. Soc., B,* **55,** 725-740.
[20] Li, K. C. (1991) Sliced Inverse Regression for dimension reduction. *J. Amer. Statist. Ass.,* **86,** 316-342.
[21] Li, K. C. (1992) On principal Hessian directions for data visualisation and dimension reduction : another application of Stein's lemma. *Ann. Statist.*, **87**, 1025-1039.
[22] Li K.C., Aragon Y., Shedden, K et Thomas-Agan C., (2003) Dimension redcution for multivariate data. *J. Amer. Statist. Ass.,* **98,** 99-109.
[23] Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis,* New-York : Springer Verlag.

[24] Wold, H. (1975). Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Square Approach. *in perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett, ed. J. Gani, London : Academic Press.*

Equipe GRIMM,, Université Toulouse le Mirail,, Toulouse,, France