Classification de données fonctionnelles et vectorielles

Boubacar Diallo, Ndeye Niang, Feriel Bouhadjera, Vincent Audigier

▶ **To cite this version:**

Boubacar Diallo, Ndeye Niang, Feriel Bouhadjera, Vincent Audigier. Classification de données fonctionnelles et vectorielles. 57èmes Journées de Statistique (JdS), Société Française de Statistique, Jun 2026, Clermond-Ferrand, France. ⟨hal-05589708⟩

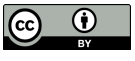
HAL Id: hal-05589708

https://hal.science/hal-05589708v1

Submitted on 13 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

CLASSIFICATION DE DONNÉES FONCTIONNELLES ET VECTORIELLES

Boubacar Diallo[1] & Ndeye Niang[1] & Feriel Bouhadjera[1] & Vincent Audigier[1]

[1] *Laboratoire CEDRIC, équipe MSDMA, 2 rue Conté, 75003 Paris*

Résumé. Nous nous intéressons à la classification d'observations décrites à la fois par une variable fonctionnelle et un ensemble de variables réelles. Une première approche directe reposant sur une distance combinant une partie fonctionnelle et une partie euclidienne est proposée. La seconde approche, appelée tandem, consiste à effectuer la classification à partir des composantes principales d'une analyse en composantes principales, dite hybride, intégrant les deux natures des données. Dans une étude par simulation, nous analysons les performances des approches via l'indice de Rand ajusté et l'indice silhouette. Les résultats mettent en évidence l'intérêt de ces deux approches lorsque les deux natures de données portent une information sur la structure en classes.

Mots-clés. Classification, données hybrides, analyse en composantes principales fonctionnelles.

Abstract. We are interested in clustering observations described by both a functional variable and a set of real variables. A first direct approach based on a distance combining a functional part and a Euclidean part is proposed. The second approach, known as tandem approach, consists in performing the clustering on the components of a principal component analysis integrating both types of data, called hybrid principal component analysis. In a simulation study, we analyze the performance of these approaches using the adjusted Rand index and the silhouette index. The results highlight the value of these two approaches when both types of data provide information about the clusters structure.

Keywords. Clustering, hybrid data functional principal component analysis.

1 Introduction

Nous nous intéressons à la classification de données dites *hybrides*, dans lesquelles chaque individu est décrit à la fois par des variables réelles (partie vectorielle) et par une variable fonctionnelle. Ce type de données apparaît dans de nombreux domaines appliqués. Par exemple, on retrouve dans Jang [2021] des travaux mettant en évidence les liens entre des courbes de rénogrammes (partie fonctionnelle) et des variables cliniques (partie vectorielle), permettant de mieux comprendre les mécanismes physiologiques sous-jacents à l'obstruction rénale.
À notre connaissance, la problématique de la classification des données hybrides n'a pas été traitée dans la littérature. On trouve cependant des travaux sur l'Analyse en Composantes Principales (ACP) de données hybrides dans Ramsay and Silverman [2005] et Jang [2021].

Pour la classification de ces données hybrides, nous proposons deux approches basées sur des méthodes géométriques classiques : k-means, PAM (*Partitionning Around Medoids*) et la Classification Ascendante Hiérarchique (CAH). Une première approche directe, repose sur une distance, dite hybride, combinant une partie euclidienne sur les variables réelles et une extension de cette dernière pour les données fonctionnelles. Une seconde approche, dite tandem, consiste à réaliser une ACP sur les données hybrides suivie d'une classification sur les composantes de cette ACP. Dans la suite, nous présenterons les deux approches directe et tandem qui seront ensuite comparées à travers une étude de simulation. Nous finirons par des conclusions et perspectives.

2 Méthodes proposées

2.1 Approche directe avec distance hybride

On considère une variable aléatoire (v.a.) dite hybride $Z = (X(t), Y)$ composée d'une v.a. fonctionnelle $X(t)$ et d'un vecteur Y de M v.a. réelles, définie dans l'espace produit $\mathcal{Z} = \mathcal{L}^2(\mathcal{T}) \times \mathbb{R}^M$, où $\mathcal{L}^2(\mathcal{T})$ est l'espace des fonctions de carrés intégrable sur le compact $\mathcal{T} \subset \mathbb{R}$. On observe un échantillon de taille n noté $(Z_i)_{1 \leq i \leq n}$ où $Z_i = (X_i(t), Y_i)$, $X_i \in \mathcal{L}^2(\mathcal{T})$ et $Y_i \in \mathbb{R}^M$. On munit l'espace $\mathcal{Z}$ du produit scalaire suivant :

$$\langle Z_1, Z_2 \rangle = \int_T X_1(t) X_2(t)\, dt + \alpha^2\, Y_1^T Y_2, \quad t \in \mathcal{T},$$

où $\alpha^2 = \dfrac{\sum_{i=1}^n \|X_i - \hat{\mu}\|^2_{\mathcal{L}^2}}{\sum_{i=1}^n \|Y_i - \bar{Y}\|^2}$ permet d'équilibrer les contributions fonctionnelle et vectorielle des n individus avec $\hat{\mu}$ qui désigne la fonction moyenne empirique et $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_M)$ le vecteur des moyennes empiriques. La distance hybride entre deux observations Z_1 et Z_2 est alors définie par :

$$\underbrace{D^2(Z_1, Z_2)}_{D_{\text{tot}}} = \langle Z_1 - Z_2, Z_1 - Z_2 \rangle = \underbrace{\int_T (X_1(t) - X_2(t))^2\, dt}_{D_0} + \alpha^2\, \underbrace{\|Y_1 - Y_2\|^2}_{D_Y}. \tag{1}$$

Cette distance est utilisée dans les méthodes géométriques de classification.

2.2 Approche tandem avec ACP hybride

Les méthodes proposées pour l'ACP hybride reposent sur une ACP classique, à l'aide des représentations vectorielles de la partie fonctionnelle. Deux approches sont distinguées dans la littérature : RS-PCA de Ramsay and Silverman [2005] et HFV-PCA, de Jang [2021].

2.2.1 RS-PCA

Dans l'approche RS-PCA, chaque courbe X_i est projetée sur une base de fonctions $\{\phi_b\}_{b=1}^B$ et représentée par son vecteur de coefficients $c_i \in \mathbb{R}^B$. On construit alors la représentation

hybride w_i en concaténant ces coefficients avec Y_i : $w_i = (c_i, \alpha Y_i) \in \mathbb{R}^{B+M}$. L'ACP hybride consiste ainsi à appliquer une ACP classique à la matrice W des représentations hybrides w_i. Pour chaque vecteur propre $u_d = (a_d, b_d) \in \mathbb{R}^B \times \mathbb{R}^M$ issu de l'ACP sur la matrice de covariance V, les coordonéess de la d-ième composante hybride sont alors les projections des w_i sur u_d :

$$\hat{\rho}_{id} = {}^t w_i u_d, \quad \text{si } \{\phi_b\}_{b=1}^B \text{ est orthonormale.}$$

Lorsque la base de lissage n'est pas orthonormale, cette représentation vectorielle de la partie fonctionnelle ne permet pas de préserver les propriétés géométriques (distance et orthogonalité).

2.2.2 HFV-PCA

Dans l'approche HFV-PCA, les deux parties sont représentées par des scores d'ACP. Une ACP fonctionnelle est appliquée aux données X_i, fournissant pour chaque individu i, des scores fonctionnels ρ_{il}, $l = 1, \ldots, L$. En parallèle, une ACP classique est appliquée à la partie vectorielle pondérée $Y_i^* = \alpha Y_i$, produisant des scores γ_{ig}, $g = 1, \ldots, G$ où α est défini tel que présenté en Section 2.1. Les nombres de composantes principales G et L sont choisis de manière à conserver respectivement une proportion δ_1 et δ_2 de la variabilité des données fonctionnelles et vectorielles. La représentation hybride de l'individu i est alors $w_i = (\rho_{i1}, \ldots, \rho_{iL}, \gamma_{i1}, \ldots, \gamma_{iG})$. Une ACP classique est ensuite appliquée à la matrice $W \in \mathbb{R}^{n \times (L+G)}$. Les vecteurs propres s'écrivent $u_d = (a_d, b_d) \in \mathbb{R}^L \times \mathbb{R}^G$ et les scores hybrides associées sont :

$$\hat{\rho}_{id} = \sum_{l=1}^{L} \rho_{il}\, a_{dl} + \sum_{g=1}^{G} \gamma_{ig}\, b_{dg}.$$

L'approche tandem proposée consiste à réaliser la classification sur les scores issus de chacune de ces ACP.

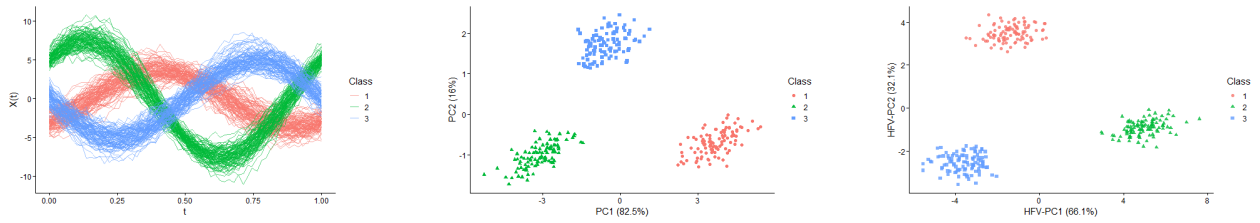
3 Expérimentations

Afin de comparer les différentes approches, deux scénarios de génération des données hybrides $Z_i = (X_i(t), Y_i)$, $t \in \mathcal{T}$, $i = 1, \ldots, n$ inspirés de Jang [2021] sont considérés. Les données comportent $K = 3$ classes, chacune constituée de 100 individus. Les données fonctionnelles $X_i(t)$ sont observées sur une grille dense et régulière de 60 points dans $\mathcal{T} = [0, 1]$.

Scénario I : Les données fonctionnelles sont générées selon le modèle : $X_i(t) = \mu^{(k)}(t) + \sum_{d=1}^{10} \rho_{id}^{(k)} \xi_d(t) + \epsilon_i$, $t \in \mathcal{T}$, où $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$. La fonction moyenne de la classe k est définie par $\mu^{(k)}(t) = \sum_{d=1}^{10} c_d^{(k)} \xi_d(t), k \in \{1, 2, 3\}$. Les coefficients $c_d^{(k)}$ sont donnés par : $c_d^{(k)} = \big(\underbrace{7, 2, \mathbf{0}_8}_{k=1}, \underbrace{0.5, 5, \mathbf{0}_8}_{k=2}, \underbrace{-5, -5, \mathbf{0}_8}_{k=3}\big)$, où $\mathbf{0}_8$ est le vecteur nul de longueur 8, de sorte que seules les deux premières dimensions séparent les classes. Les fonctions propres $\{\xi_d\}_{d=1}^{10}$ correspondent aux 10 premières fonctions d'une base de Fourier linéairement indépendantes et orthogonales

sur $\mathcal{T}$. Les coefficients de décomposition associés sont simulés selon $\rho_{id}^{(k)} \sim \mathcal{N}(m^{(k)}, \lambda_d^{(k)})$ avec $m^{(k)} = \big(\underbrace{-5,-5,\mathbf{0}_8}_{k=1}, \underbrace{5,0,\mathbf{0}_8}_{k=2}, \underbrace{0,5,\mathbf{0}_8}_{k=3}\big)$ et $\lambda_d^{(k)} = \big(\underbrace{0.5^{d-1}}_{k=1}, \underbrace{0.4^{d-1}}_{k=2}, \underbrace{0.45^{d-1}}_{k=3}\big)$ pour $d = 1, \ldots, 10$.

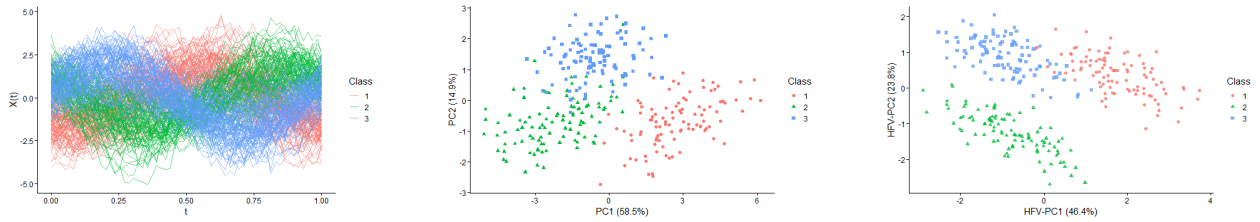
La décroissance rapide de $\lambda_d^{(k)}$ en fonction de d assure une bonne représentation des données sur les deux premières composantes principales fonctionnelles. Les données vectorielles sont quant à elles générées selon : $Y_i = \sum_{d=1}^{10} \rho_{id}^{(k)} \theta_d$, où $\{\theta_d\}_{d=1}^{10}$ sont les vecteurs propres de la décomposition d'une matrice de corrélation $R \in \mathbb{R}^{10\times 10}$ telle que $R_{d_1 d_1} = 1$ et $R_{d_1 d_2} = 0.2$ pour $d_1 \neq d_2$ avec $d_1, d_2 = 1, \ldots, 10$. Ainsi, le lien entre les données vectorielles et fonctionnelles est gérée via les coefficients $\rho_{id}^{(k)}$, identiques dans les deux cas.
La Figure 1 présente les données générées selon le scénario I. On observe que les courbes sont clairement distinctes et les classes sont bien séparées, aussi bien sur les deux premières composantes principales d'une ACP appliquée à la partie vectorielle que via l'ACP hybride avec HFV-PCA.



(a) (b) (c)

FIGURE 1 – Données générées selon le scénario I. (a) Courbes lissées, (b) ACP sur Y et (c) HFV-PCA.

Scénario II : Dans ce second scénario, la séparation entre les classes est atténuée. Les coefficients des fonctions moyennes $c_d^{(k)}$ et les moyennes $m^{(k)}$ des coefficients $\rho_{id}^{(k)}$ sont définis par $c_d^{(k)} = \big(1.5, 0.5, \mathbf{0}_8,\ -3, 1, \mathbf{0}_8,\ 1.5, -1.5, \mathbf{0}_8\big)$ et $m^{(k)} = \big(-2, -2, \mathbf{0}_8,\ 2, 0, \mathbf{0}_8,\ 0, 2, \mathbf{0}_8\big)$.
Sur la partie vectorielle, on ajoute un bruit gaussien indépendant $\eta_i \sim \mathcal{N}(0, 0.2^2)$ à Y_i. Contrairement au scénario I, où l'indice silhouette est proche de 1, il est ici d'environ 0.4 (voir Table 1) traduisant une structure en classes peu marquée. En effet, dans la Figure 2 (a), les courbes apparaissent plus mélangées. De plus, on observe des Figures 2 (b) et (c) que la séparation en classes est moins marquée sur les deux premières composantes principales de l'ACP appliquée à la partie Y ou de l'ACP hybride avec HFV-PCA.



(a) (b) (c)

FIGURE 2 – Données générées selon le scénario II. (a) Courbes lissées, (b) ACP sur Y et (c) HFV-PCA.

Méthodes de classification : les données fonctionnelles étant observées sur une grille discrète, elles sont d'abord lissées par moindres carrés pénalisés à l'aide de 30 fonctions de base de Fourier. Le paramètre de pénalisation est choisi par validation croisée généralisée. Pour l'approche directe, après lissage, les méthodes (k-means, PAM et une CAH avec une stratégie de Ward, consolidée) sont appliquées à la matrice des distances calculée à partir de (1). L'intégrale intervenant dans la distance est approchée par une somme de Riemann. Pour l'approche tandem (RS-PCA et HFV-PCA), un nombre suffisant de composantes principales est retenu afin d'expliquer au moins 99% de la variabilité des données ($\delta_1 = \delta_2 = 0.99$).

Critères d'évaluation : Les partitions obtenues sont comparées à la partition de référence à l'aide de l'indice de Rand ajusté (ARI). Les résultats sont présentés sous forme de moyenne et d'écart-type sur 100 répétitions pour chacun des deux scénarios. Les indices de silhouette correspondants sont également fournis, les distances étant définies par l'équation (1).

3.1 Résultats

Approche tandem vs directe : Dans le scénario I, toutes les méthodes retrouvent parfaitement la partition de référence (ARI = 1). Cela s'explique par le fait que les classes sont bien séparées sur les deux parties, comme l'indiquent les indices de silhouette élevés (compris entre 0.778 et 0.808, voir Table 1 et 3.1).

Dans le scénario II, l'indice silhouette est faible (moins de 0.4). Les résultats montrent qu'avec k-means et la CAH consolidée, l'approche directe est meilleure (ARI de $D_{\text{tot}} = 0.960$ contre 0.926 pour RS-PCA et 0.927 pour HFV-PCA). Cependant, avec PAM, on observe le contraire (0.910 pour RS-PCA et 0.906 pour HFV-PCA contre 0.833 pour D_{tot}).

Données hybrides vs données fonctionnelles ou vectorielles : Dans le scénario I, la forte séparation des classes ne permet pas de mettre en évidence des différences entre les approches. En revanche, dans le scénario II, pour toutes les méthodes de classification, la prise en compte conjointe des données fonctionnelles et vectorielles améliore la capacité à retrouver la partition de référence : à titre d'exemple, avec k-means, l'ARI est de 0.960 (D_{tot}) lorsque les deux parties sont combinées, de 0.701 pour la partie fonctionnelle (D_0) et de 0.831 pour la partie vectorielle (D_Y). Ce dernier résultat est cohérent avec $\alpha^2 \approx 0.225$ (Table 2) qui indique plus de variabilité dans cette partie vectorielle. Cela met en évidence l'apport des approches hybrides lorsque la structure des classes est moins nette.

Scénarios	D_0	D_Y	D_{tot}
I	0.778 (0.006)	0.783 (0.006)	0.808 (0.005)
II	0.331 (0.022)	0.325 (0.014)	0.361 (0.015)

TABLE 1 – Indices de silhouette de la vraie partition

Scénarios	Moyenne	Écart-type
I	0.450	0.009
II	0.225	0.011

TABLE 2 – Statistiques de α^2

Scénarios	Approches	k-means		PAM		CAH	
		SIL	ARI	SIL	ARI	SIL	ARI
I	Tandem RS-PCA	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)
	Tandem HFV-PCA	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)
	Directe D_{tot}	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)	**0.808** (0.005)	**1.000** (0.000)
	Directe D_0	0.778 (0.006)	1.000 (0.001)	0.778 (0.006)	1.000 (0.001)	0.778 (0.006)	1.000 (0.001)
	Directe D_Y	0.783 (0.006)	1.000 (0.000)	0.783 (0.006)	1.000 (0.000)	0.783 (0.006)	1.000 (0.000)
II	Tandem RS-PCA	**0.366** (0.014)	**0.926** (0.026)	**0.364** (0.014)	**0.910** (0.039)	**0.366** (0.014)	**0.926** (0.025)
	Tandem HFV-PCA	**0.366** (0.014)	**0.927** (0.026)	**0.364** (0.015)	**0.906** (0.045)	**0.366** (0.014)	**0.927** (0.025)
	Directe D_{tot}	**0.367** (0.014)	**0.960** (0.019)	**0.371** (0.014)	**0.833** (0.044)	**0.367** (0.014)	**0.960** (0.019)
	Directe D_0	0.396 (0.014)	0.701 (0.049)	0.392 (0.015)	0.685 (0.054)	0.396 (0.014)	0.701 (0.049)
	Directe D_Y	0.347 (0.011)	0.831 (0.040)	0.340 (0.014)	0.793 (0.054)	0.347 (0.011)	0.831 (0.040)

TABLE 3 – Performance des méthodes de classification dans les deux scénarios. En gras, lorsque les données hybrides sont considérées.

4 Conclusions et perspectives

Dans ce travail, nous nous sommes intéressés à la classification de données hybrides combinant une partie fonctionnelle et une partie vectorielle. À travers une étude de simulation en deux scénarios, nous avons montré l'intérêt des approches hybrides lorsque les deux parties portent une information sur la structure en classes. Les bons résultats observés pour des classes bien séparées sont plus mitigés lorsque la séparation en classe est moins nette dans les deux parties, en particulier pour l'approche tandem avec l'ACP hybride. En perspectives, il sera nécessaire de valider ces conclusions sur des données plus complexes (grilles éparses sur la partie fonctionnelle, plusieurs variables fonctionnelles). De plus, l'intégration de la dérivée première des fonctions, rajoutant une discrimination suivant la forme [Diallo et al., 2026], pourrait améliorer les performances des méthodes proposées.

Bibliographie

B. Diallo, N. Niang, V. Audigier, and F. Bouhadjera. Comparaison de méthodes de classification de données fonctionnelles. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-42 :121–132, 2026.

J. H. Jang. Principal component analysis of hybrid functional and vector data. *Statistics in medicine*, 40(24) :5152–5173, 2021.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2 edition, 2005.