Une approche par noyaux multiples pour l'apprentissage non-supervisé de représentation de données fonctionnelles dans des espaces de Sobolev

Julien Ah-Pine[1,2] & Anne-Françoise Yao[2]

[1] *Université de Lyon, Lyon 2 et ERIC EA3083, 5 Avenue Pierre Mendèes France, 69500 Bron, France; julien.ah-pine@univ-lyon2.fr*

[2] *Université Clermont Auvergne, LMBP UMR6620, 3 place Vasarely, 63170 Aubière, France; anne.yao@uca.fr*

Résumé. Nous appliquons de façon complémentaire l'ACP à noyaux et le k-means à noyaux multiples aux données fonctionnelles. Nous définissons ainsi une approche pour l'apprentissage non-supervisé de ce type de données. Nous supposons que les fonctions appartiennent à un espace de Sobolev et exploitons les fonctions dérivées dans l'analyse selon une approche multi-vue. Notre méthode met en avant l'utilisation de fonctions noyaux permettant de représenter les données fonctionnelles dans des RKHS. Par ailleurs, elle utilise le k-means à noyaux multiples afin de déterminer une combinaison linéaire des fonctions noyaux qui optimise la variance inter-groupe. L'ACP à noyaux permet de réduire en amont les RKHS aux axes principaux et/ou de visualiser, en aval, les résultats du k-means à noyaux multiples.

Mots-clés. Analyse de données fonctionnelles, Fonctions dérivées, k-means à noyaux multiples, ACP à noyaux.

Abstract. We apply kernel PCA and multiple kernel k-means to functional data in a complementary way. We introduce a framework for the unsupervised learning of that kind of data. We assume that the functions belong to a Sobolev space and we emphasize the derivative functions in the analysis following a multi-view approach. Our framework makes it possible to use kernel functions in order to implicitly represent the functions and their derivatives in RKHS. Moreover, it uses a multiple kernel k-means so as to determine a linear combination of the kernel functions that aims at maximizing the variance between the clusters. The kernel PCA can be used to reduce each kernel matrix before clustering and/or to visualize the results of the multiple kernel k-means.

Keywords. Functional data analysis, Derivative functions, multiple kernel k-means, kernel PCA.

1 Contexte et description de l'approche proposée

Les technologies modernes nous permettent d'enregistrer de façon volumineuse des mesures de phénomènes évoluant dans le temps et l'espace. Ces phénomènes peuvent être très

divers allant du domaine de l'environnement au niveau planétaire comme pour le réchauffement climatique, jusqu'aux activités quotidiennnes de chaque individu comme la mesure de la fréquence cardiaque tout au long d'une journée.

Du point de vue formel, ces ensembles discrets de mesures proviennent de fonctions continues que l'on observe en des points dans le temps et/ou l'espace. Ce type de données dépassent le cadre classique multivarié puisque ce dernier ne tient pas compte explicitement de cette dépendance temporelle et/ou spatiale. Or pour l'analyse des phénomènes sous-jacents, il est important d'intégrer ces dépendances comme composante essentielle de la nature même des données que l'on étudie. L'analyse de données fonctionnelles est la branche des statistiques qui s'intéresse à cette problématique [8, 2].

Dans cette communication, les objets à l'étude forment un échantillon de n fonctions $\{x_i\}_{i=1,\dots,n}$. Nous supposons que ce sont des éléments de l'espace de Sobolev $\mathbb{W}^{2,q}$ consistant en des fonctions de $\mathbb{L}^2$ dont les dérivées $\{D^j x_i\}_{i=1,\dots,n;j=1,\dots,q}$ sont également des fonctions de $\mathbb{L}^2$. Nous nous intéressons plus particulièrement au problème de l'apprentissage non-supervisé où il s'agit de déterminer les principales régularités au sein de l'échantillon. Deux applications distinctes mais complémentaires dans ce cadre sont: la réduction de dimension et la classification automatique.

L'approche classique pour la réduction de dimension est l'analyse en composantes principales (ACP) fonctionnelle. Celle-ci repose sur l'analyse spectrale de l'opérateur covariance. En particulier, la fonction covariance est un noyau de Mercer et on peut donc la décomposer dans une base orthonormée de fonctions propres. L'ACP fonctionnelle projette alors les $\{x_i\}_i$ dans le sous-espace engendré par les fonctions propres associées aux valeurs propres les plus grandes afin de conserver au maximum la variance.

Nous proposons d'appliquer une démarche duale qui consiste non pas à étudier l'opérateur de covariance mais la matrice de Gram associée à l'échantillon. En particulier, nous utilisons ici des fonctions noyaux permettant de représenter implicitement les fonctions dans des espaces de Hilbert à noyau reproduisant (RKHS). Nous notons par $\mathbf{K}$ la matrice de Gram de taille $(n \times n)$ et de terme général $\mathbf{K}_{i,i'} = k(x_i, x_{i'})$ où $k : \mathbb{L}^2 \times \mathbb{L}^2 \to \mathbb{R}$ est une fonction noyau symétrique et définie positive. La motivation de cette démarche est similaire à celle au cas multivarié: les $\{x_i\}_i$ peuvent appartenir à des sous-espaces non linéaires de $\mathbb{L}^2$ et dans ce cas leur représentation dans des RKHS pourrait permettre de mieux appréhender cette non-linéarité.

Nous proposons ainsi d'utiliser l'ACP à noyaux [10] pour l'étude de données fonctionnelles. Notons que dans ce cas, il n'est pas nécessaire d'appliquer quelconque modification à la méthode définie dans le cadre multivarié dans la mesure où la structure algébrique de base qu'elle utilise est un espace de Hilbert séparable ce qui est notre cas ici également. Ceci fût déjà discuté dans [9] qui étend les SVM aux données fonctionnelles.

En revanche, nous mettons en avant la nature fonctionnelle des données et supposons en particulier que les fonctions appartiennent à $\mathbb{W}^{2,q}$. Notre hypothèse, classique en analyse de données fonctionnelles, est que les fonctions dérivées peuvent apporter une information pertinente voire cruciale pour l'analyse.

Nous représentons donc un élément x_i de notre échantillon, par la fonction elle-même mais également par ses fonctions dérivées: $(x_i, D^1x_i, \ldots, D^qx_i)$. Ces $q+1$ fonctions sont dans $\mathbb{L}^2$ mais l'utilisation de fonctions noyaux nous permettent également de les représenter dans des RKHS $\{\mathbb{H}^j\}_{j=0,\ldots,q}$ et dans ce cas, on suppose qu'il existe $q+1$ applications $\{\phi^j : \mathbb{L}^2 \to \mathbb{H}^j\}_{j=0,\ldots,q}$ nous permettant d'étudier implicitement et de façon plus large $(\phi^0(x_i), \phi^1(D^1x_i), \ldots, \phi^q(D^qx_i))$. Cette approche permet ainsi un cadre riche pour la représentation des données fonctionnelles en apprentissage automatique.

Il n'en reste pas moins la question de la métrique qui serait la plus avantageuse pour étudier les relations de proximité entre ces fonctions. Dans cette perspective nous supposons le modèle suivant: $\langle x_i, x_{i'}\rangle = \sum_{j=0}^{q} w_j \langle \phi^j(D^jx_i), \phi^j(D^jx_{i'})\rangle_{\mathbb{H}^j}$ où $w_j \geq 0, \forall j = 0, \ldots, q$, pour que la combinaison linéaire donne un noyau valide. Notons que, en appliquant l'astuce du noyau, ceci est équivalent à se donner $q+1$ fonctions noyaux $\{k^j\}_{j=1,\ldots,q}$ et dans ce cas le modèle s'écrit comme suit:

$$k(x_i, x_{i'}) = \sum_{j=0}^{q} w_j k^j(D^jx_i, D^jx_{i'})$$

Notons $\{\mathbf{K}^j\}_{j=0,\ldots,q}$ les $q+1$ matrices de Gram de taille $(n \times n)$. Chaque matrice de Gram peut-être interprétée telle une vue distincte du même élément. Notre problème consiste alors à déterminer $\mathbf{w} = (w_j)_{j=0,\ldots,q}$ tel que $\mathbf{K} = \sum_{j=0}^{q} w_j \mathbf{K}^j$ donne une matrice de Gram efficace pour l'apprentissage non-supervisé c'est à dire qui contribue à faire ressortir à la fois les proximités des éléments formant un groupe homogène et les disparités entre les éléments appartenant à des groupes distincts.

Dans ce contexte, nous soulignons l'importance de standardiser les matrices de Gram afin qu'elles soient mutuellement commensurables. Pour cela, nous divisons chaque matrice de Gram par l'écart-type des valeurs qu'elle contient. De plus, nous proposons d'utiliser une méthode basée sur les k-means à noyaux multiples afin d'estimer $\mathbf{w}$. La méthode consiste à maximiser la variance inter-groupe qui dépend ici de deux variables: $\mathbf{P}$ la partition des éléments en k groupes et $\mathbf{w}$ le vecteur des poids des matrices de Gram donnant le noyau agrégé. Dans les travaux précédents en k-means à noyaux multiples (voir par exemple [11]), plusieurs types de contraintes ont été imposées au vecteur $\mathbf{w}$ afin de borner le problème. Il est à noter que la contrainte $\sum_j w_j = 1$ donne en sortie un vecteur sparse. Dans notre cas, nous choisissons d'utiliser la contrainte $\sum_j w_j^2 = 1$ qui permet également une solution analytique mais qui aboutit à un véritable mélange des matrices de Gram.

Notre approche de k-means à noyaux multiples appliquée aux données fonctionnelles aboutit à deux types de résultats: d'une part, nous obtenons une partition de l'échantillon qui repose sur l'information provenant des fonctions et de leurs dérivées (représentées dans des RKHS ou pas); d'autre part, nous apprenons une combinaison linéaire qui combine les différentes matrices de Gram donnant ainsi un noyau k optimisé.

Enfin, nous proposons d'appliquer l'ACP à noyau sur la matrice de Gram $\mathbf{K}$ issue de la combinaison optimale afin de visualiser l'échantillon dans un espace réduit. Nous

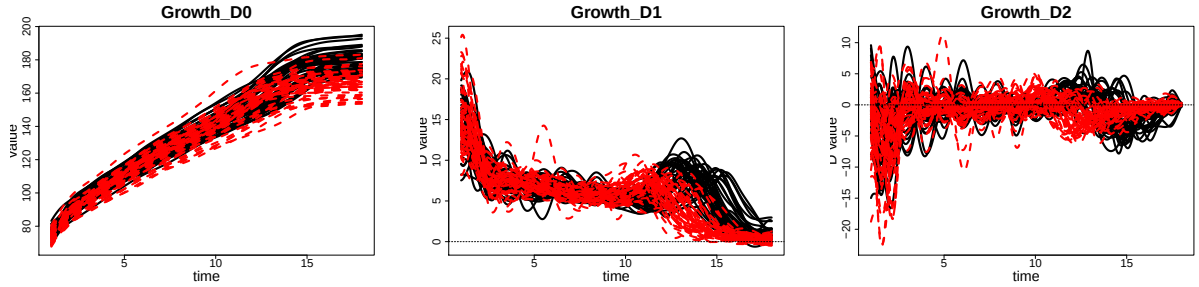

Figure 1: Courbes du jeu de données Growth. De gauche à droite: $\{x_i\}_i$, $\{Dx_i\}_i$, $\{D^2x_i\}_i$. En rouge les filles et en noir les garçons.

montrons également comment il est possible de calculer les contributions des différentes vues à la constitution des axes ce qui permet de mieux apprécier les sources principales de discrimination que notre approche a mise en avant.

2 Illustration des résultats de l'approche proposée

Nous illustrons les résultats de notre méthode à l'aide du jeu de données Growth de Berkeley qui correspond à des mesures de la taille de 93 enfants à plusieurs moments de leur première partie de vie. Nous cherchons à vérifier s'il est facilement possible de différencier de façon non-supervisée les courbes de croissance des filles de celles des garçons.

Dans la Figure 1 nous montrons les courbes des 93 sujets en distinguant en rouge les courbes des filles et en noir celles des garçons. A partir des données brutes qui sont des observations discrètes de chaque courbe, nous reconstituons la forme fonctionnelle des éléments, c'est à dire les $\{x_i\}_i$, en les représentant dans une base de fonctions B-splines. Les fonctions dérivées premières et secondes $\{Dx_i\}_i$ et $\{D^2x_i\}_i$ sont également déterminées dans cette base.

Dans la Figure 2, nous montrons les résultats de notre approche. Nous confrontons la partition obtenue par la méthode k-means à la vérité terrain par le biais de la mesure de l'information mutuelle normalisée[1] (NMI). Cinq représentations sont testées $\{x_i\}_i$, $\{Dx_i\}_i$, $\{D^2x_i\}_i$, $\{(x_i, Dx_i)\}_i$ et $\{(x_i, Dx_i, D^2x_i)\}_i$. Dans le cas des deux dernières, nous utilisons un k-means à noyaux multiples que l'on choisit de même nature:

$$k(x_i, x_i') = \sum_{j=0}^{q} w_j k'(D^j x_i, D^j x_{i'})$$

[1]Plus la mesure est proche de 1, meilleur est le résultat de la classification automatique.

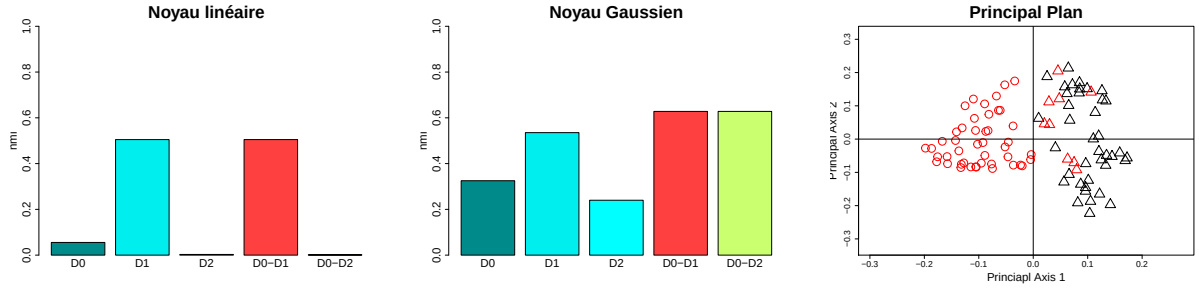

Figure 2: Résultats obtenus par notre approche. Le diagramme de gauche montre les mesures NMI pour les cinq représentations avec un noyau linéaire. Le diagramme du centre correspond aux performances pour le noyau Gaussien. Le nuage de points de droite est le résultat de l'ACP à noyau appliqué au noyau agrégé issu des noyaux Gaussiens avec la représentation $\{(x_i, Dx_i, D^2x_i)\}_i$.

avec $q = 1$ ou $q = 2$ et k' est le noyau linéaire ou (exclusif) Gaussien[2].

Le diagramme en bâtons à gauche de la Figure 2, montre les performances du k-means avec le noyau linéaire. Si nous utilisons une seule vue, clairement, c'est la dérivée première qui donne les meilleurs résultats. Combiner $\{(x_i, Dx_i)\}_i$ donne d'aussi bons résultats mais ce n'est pas le cas pour $\{(x_i, Dx_i, D^2x_i)\}_i$. En effet, malgré l'optimisation du vecteur poids, il semble que la dérivée seconde dans la représentation linéaire, apporte un bruit important qui conduit à un mauvais noyau agrégé.

Le diagramme en bâtons au centre expose les résultats obtenus avec un noyau Gaussien. Il est intéressant de noter que l'utilisation d'une fonction noyau non-linéaire permet ici d'améliorer les performances de chaque vue. Par ailleurs, nous obtenons cette fois-ci de très bons résultats pour le k-means à noyaux multiples en combinant les matrices de Gram des fonctions avec celles des dérivées premières (bâton rouge) puis celles des dérivées secondes (bâton vert).

Le nuage de points à droite de la Figure 2 correspond à la projection des éléments de l'échantillon obtenue par l'ACP à noyau. La matrice de Gram utilisée dans ce cas, est celle obtenue par le k-means à noyaux Gaussiens multiples avec la représentation $\{x_i, Dx_i, D^2x_i\}_i$. La partition à 2 classes qui est solution du k-means à noyaux multiples, peut être visualisée par les symboles cercles *versus* triangles. La vérité terrain est représentée par les couleurs. Ainsi les triangles en rouge à droite du plan sont les erreurs de notre approche non-supervisée (9 sur 93, soit un taux d'erreur de moins de 10%).

[2]Dans ce cas, nous fixons le paramètre σ^2 égale à la médiane des distances $\mathbb{L}^2$ au carré de chaque courbe à son 7ème plus proche voisin. Ce paramétrage est motivé par les résultats empiriques exposés dans [15].

Bibliographie

[1] C. Abraham, P.-A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, 30(3):581–595, 2003.

[2] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[3] D. Floriello and V. Vitelli. Sparse clustering of functional data. *Journal of Multivariate Analysis*, 154:1–18, 2017.

[4] M. L. L. García, R. García-Ródenas, and A. G. Gómez. K-means algorithms for functional data. *Neurocomputing*, 151:231–245, 2015.

[5] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[6] Y. Meng, J. Liang, F. Cao, and Y. He. A new distance with derivative information for functional k-means clustering algorithm. *Information Sciences*, 463:166–185, 2018.

[7] A. Muñoz and J. González. Representing functional data using support vector machines. *Pattern Recognition Letters*, 31(6):511–516, 2010.

[8] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

[9] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730–742, 2006.

[10] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[11] G. Tzortzis and A. Likas. Kernel-based weighted multi-view clustering. In *2012 IEEE 12th international conference on data mining*, pages 675–684. IEEE, 2012.

[12] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[13] D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[14] M. Yamamoto. Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, 6(3):219–247, 2012.

[15] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.